Scalable Online Exploration via Coverability

Philip Amortila [*1] **Dylan J. Foster** [*2] **Akshay Krishnamurthy** [*2]

Abstract

Exploration is a major challenge in reinforcement learning, especially for high-dimensional domains that require function approximation. We propose *exploration objectives*—policy optimization objectives that enable downstream maximization of any reward function—as a conceptual framework to systematize the study of exploration. We introduce a new objective, L_1-*Coverage*, which generalizes previous exploration schemes and supports three fundamental desiderata:

1. *Intrinsic complexity control.* L_1-Coverage is associated with a structural parameter, L_1-*Coverability*, which reflects the intrinsic statistical difficulty of the underlying MDP, subsuming Block and Low-Rank MDPs.
2. *Efficient planning.* For a known MDP, optimizing L_1-Coverage efficiently reduces to standard policy optimization, allowing flexible integration with off-the-shelf methods such as policy gradient and Q-learning approaches.
3. *Efficient exploration.* L_1-Coverage enables the first computationally efficient model-based and model-free algorithms for online (reward-free or reward-driven) reinforcement learning in MDPs with low coverability.

Empirically, we find that L_1-Coverage effectively drives off-the-shelf policy optimization algorithms to explore the state space.

1. Introduction

Many applications of reinforcement learning and control demand agents to maneuver through complex environments with high-dimensional state spaces that necessitate function approximation and sophisticated exploration. Toward addressing the high sample complexity of existing empirical paradigms (Mnih et al., 2015; Silver et al., 2016; Kober et al., 2013; Lillicrap et al., 2015; Li et al., 2016), a recent body of theoretical research provides structural conditions that facilitate sample-efficient exploration, as well as an understanding of fundamental limits (Russo & Van Roy, 2013; Jiang et al., 2017; Wang et al., 2020b; Du et al., 2021; Jin et al., 2021a; Foster et al., 2021; 2023). Yet, computational efficiency remains a barrier: outside of simple settings (Azar et al., 2017; Jin et al., 2020b), our understanding of algorithmic primitives for efficient exploration is limited.

In this paper, we propose *exploration objectives* as a conceptual framework to develop efficient algorithms for exploration. Informally, an exploration objective is an optimization objective that incentivizes a policy (or policy ensemble) to explore the state space and gather data that can be used for downstream tasks (e.g., policy optimization or evaluation). To enable practical and efficient exploration, such an objective should satisfy three desiderata:

1. *Intrinsic complexity control.* Any policy (ensemble) that optimizes the objective should cover the state space to the best extent possible, enabling sample complexity guarantees for downstream learning that reflect the intrinsic statistical difficulty of the underlying MDP.
2. *Efficient planning.* When the MDP of interest is *known*, it should be possible to optimize the objective in a computationally efficient manner, ideally in a way that flexibly integrates with existing pipelines.
3. *Efficient exploration.* When the MDP is *unknown*, it should be possible to optimize the objective in a computationally *and* statistically efficient manner; the first two desiderata are necessary, but not sufficient here.

Our development of exploration objectives, particularly our emphasis on integrating with existing pipelines, is motivated by the large body of empirical research equipping policy gradient methods and value-based methods with exploration bonuses (Bellemare et al., 2016; Tang et al., 2017; Pathak et al., 2017; Martin et al., 2017; Burda et al., 2018; Ash et al., 2022). Although a number of prior theoretical works either implicitly or explicitly develop exploration objectives, they are either too general to admit efficient planning and exploration (Jiang et al., 2017; Dann et al., 2018; Jin et al.,

[1]University of Illinois, Urbana-Champaign. [2]Microsoft Research. Full version appears at [arXiv:2403.06571].

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

2021a; Foster et al., 2021; Chen et al., 2022b; Xie et al., 2023; Liu et al., 2023b), or too narrow to apply to practical problems of interest without losing fidelity to the theoretical foundations (e.g., Azar et al., 2017; Jin et al., 2018; Dani et al., 2008; Li et al., 2010; Wagenmaker & Jamieson, 2022). The difficulty is that designing exploration objectives is intimately tied to understanding what makes an MDP easy or hard to explore, a deep statistical problem. A useful objective must succinctly distill such understanding into a usable, operational form, but finding the right sweet spot between generality and tractability is challenging. We believe our approach and our results strike this balance.

Contributions. We introduce a new exploration objective, L_1-Coverage, which flexibly supports computationally and statistically efficient exploration, satisfying desiderata (1), (2), and (3). L_1-Coverage is associated with an intrinsic structural parameter, L_1-*Coverability* (Xie et al., 2023; Amortila et al., 2024), which controls the sample complexity of reinforcement learning in nonlinear function approximation settings, subsuming Block and Low-Rank MDPs. We prove that data gathered with a policy ensemble optimizing L_1-Coverage supports downstream policy optimization with general function approximation (Appendix D).

For planning in a known MDP, L_1-Coverage can be optimized efficiently through reduction to (reward-driven) policy optimization, allowing for integration with off-the-shelf methods such as policy gradient (e.g., PPO) or Q-learning (e.g., DQN). For online reinforcement learning, L_1-Coverage yields the first computationally and statistically efficient model-based and model-free algorithms for (reward-free/driven) exploration in L_1-coverable MDPs. Technically, these results can be viewed as a successful algorithmic application of the *Decision-Estimation Coefficient (DEC)* framework of Foster et al. (2021; 2023).

We complement these theoretical results with an empirical validation, where we find that the L_1-Coverage objective effectively integrates with off-the-shelf policy optimization algorithms, augmenting them with the ability to explore the state space widely.

Paper organization. In what follows, we formalize exploration objectives (Section 2), then introduce the L_1-Coverage objective (Section 3) and give algorithms for efficient planning (Section 4) and model-based exploration (Section 5), along with experiments (Section 6).

2. Online Reinforcement Learning and Exploration Objectives

This paper focuses on *reward-free reinforcement learning* (Jin et al., 2020a; Wang et al., 2020a; Chen et al., 2022b), in which the aim is to compute an exploratory ensemble of policies that enables optimization of any downstream reward function; we consider planning (computing exploratory policies in a known MDP) and online exploration (discovering exploratory policies in an unknown MDP).

We work in an episodic finite-horizon setting. With H denoting the horizon, a (reward-free) Markov decision process $M = \{\mathcal{X}, \mathcal{A}, \{P_h^M\}_{h=0}^H\}$, consists of a state space $\mathcal{X}$, an action space $\mathcal{A}$, a transition distribution $P_h^M : \mathcal{X}\times\mathcal{A} \to \Delta(\mathcal{X})$, with the convention that $P_0^M(\cdot \mid \varnothing)$ is the initial state distribution. An episode in the MDP M proceeds according to the following protocol. At the beginning of the episode, the learner selects a randomized, non-stationary *policy* $\pi = (\pi_1, \ldots, \pi_H)$, where $\pi_h : \mathcal{X} \to \Delta(\mathcal{A})$; we let Π_{rns} denote the set of all such policies, and Π_{ns} denote the subset of deterministic policies. The episode evolves through the following process, beginning from $x_1 \sim P_0^M(\cdot \mid \varnothing)$. For $h = 1, \ldots, H$: $a_h \sim \pi_h(x_h)$ and $x_{h+1} \sim P_h^M(\cdot \mid x_h, a_h)$. We let $\mathbb{P}^{M,\pi}$ denote the law under this process, and let $\mathbb{E}^{M,\pi}$ denote the corresponding expectation.

For *planning*, where the underlying MDP is known, we denote it by M. For *online exploration*, where the MDP is unknown, we denote it by $M^\star$; in this framework, the learner must explore by interacting with $M^\star$ in a sequence of *episodes* in which they execute a policy π and observe the trajectory $(x_1, a_1), \ldots, (x_H, a_H)$ that results.

Additional notation. To simplify presentation, we assume that $\mathcal{X}$ and $\mathcal{A}$ are countable; our results extend to handle continuous variables with an appropriate measure-theoretic treatment. For an MDP M and policy π, we define the induced *occupancy measure* for layer h via $d_h^{M,\pi}(x, a) = \mathbb{P}^{M,\pi}[x_h = x, a_h = a]$ and $d_h^{M,\pi}(x) = \mathbb{P}^{M,\pi}[x_h = x]$. We use π_{unif} to denote the uniform policy. We define $\pi \circ_h \pi'$ as the policy that follows π for layers $h' < h$ and follows π' for $h' \geq h$. For a set $\mathcal{Z}$, we let $\Delta(\mathcal{Z})$ denote the set of all probability distributions over $\mathcal{Z}$. We write $f = \widetilde{O}(g)$ to denote that $f = O(g \cdot \max\{1, \mathrm{polylog}(g)\})$.

2.1. Exploration Objectives

We introduce *exploration objectives* as a conceptual framework to study exploration in reinforcement learning. Exploration objectives are policy optimization objectives; they are defined over ensembles of policies (*policy covers*), represented as distributions $p \in \Delta(\Pi)$ for a class $\Pi \subset \Pi_{\mathsf{rns}}$ of interest. The defining property of an exploration objective is to incentivize policy ensembles $p \in \Delta(\Pi)$ to explore the state space and gather data that can be used for downstream policy optimization or evaluation (i.e., offline RL).

Definition 2.1 (Exploration objective)**.** *For a reward-free MDP M and policy class Π, a function $\Phi^M : \Delta(\Pi) \to \mathbb{R}_+$ is an* exploration objective *if for any policy ensemble $p \in \Delta(\Pi)$, one can optimize any downstream reward*

function R to precision ϵ in M (i.e., produce $\widehat{\pi}$ such that $\max_\pi J_R^M(\pi) - J_R^M(\widehat{\pi}) \leq \epsilon$)[1] *using* $\mathrm{poly}(\Phi^M(p), \epsilon^{-1})$ *trajectories drawn from $\pi \sim p$ (under standard function approximation assumptions).*

Note that we allow the exploration objective to depend on the underlying MDP M; if M is unknown, then evaluating $\Phi^M(p)$ may be impossible without first exploring. We also consider *per-layer* objectives denoted as Φ_h^M, optimized by a collection $\{p_h\}_{h=1}^H$ of policy ensembles. We deliberately leave the details for reward optimization vague in Definition 2.1, as this will depend on the policy optimization and function approximation techniques under consideration; see Appendix D for examples.

With Definition 2.1 in mind, the desiderata in Section 1 can be restated formally as follows.

1. *Intrinsic complexity control.* The optimal objective value $\mathsf{Opt}^M := \inf_{p\in\Delta(\Pi)} \Phi^M(p)$ is bounded for a large class of MDPs M of interest, ideally in a way that depends on intrinsic structural properties of the MDP.

2. *Efficient planning.* When the MDP M is known, we can solve $\arg\min_{p\in\Delta(\Pi)} \Phi^M(p)$ approximately (up to multiplicative or additive approximation factors) in a computationally efficient fashion, ideally in a way that reduces to standard computational primitives such as reward-driven policy optimization.

3. *Efficient exploration.* When the MDP M is unknown, we can approximately solve $\arg\min_{p\in\Delta(\Pi)} \Phi^M(p)$ in a sample-efficient fashion, with sample complexity polynomial in the optimal objective value Opt^M, and with computational efficiency comparable to planning.

As a basic example, for tabular MDPs where $|\mathcal{X}|$ is small, the work of Jin et al. (2020a) can be viewed as optimizing the per-layer objective

$$\Phi_h^M(p) = \max_{x\in\mathcal{X}} \frac{1}{\mathbb{E}_{\pi\sim p}[d_h^\pi(x)]} \tag{1}$$

for each layer $h \in [H]$; the optimal value for this objective satisfies $\mathsf{Opt}_h^M = O(|\mathcal{X}|/\eta)$, where $\eta := \min_{x\in\mathcal{X}} \max_{\pi\in\Pi_{\mathsf{rns}}} d_h^\pi(x)$ is a *reachability parameter*. This objective supports efficient planning and exploration, satisfying desiderata (2) and (3), but is restrictive in terms of intrinsic complexity (desideratum (1)) because the optimal value, which scales with $|\mathcal{X}|$, does not sharply control the sample complexity of reinforcement learning in the function approximation regime. Other (implicit or explicit) objectives studied in prior work similarly do not satisfy desideratum (1) (Hazan et al., 2019; Jin et al., 2020b; Agarwal et al., 2020; Modi et al., 2024; Uehara et al., 2022; Zhang et al., 2022; Mhammedi et al., 2023b), or are too general to admit computationally efficient planning and exploration, failing to satisfy desiderata (2) and (3) (Jiang et al., 2017; Dann et al., 2018; Jin et al., 2021a; Foster et al., 2021; Chen et al., 2022b; Xie et al., 2023; Liu et al., 2023b). See Appendix A for further discussion.

Remark 2.1. Many existing algorithms—for example, those that use count-based bonuses (Azar et al., 2017; Jin et al., 2018) or elliptic bonuses (Auer, 2002; Dani et al., 2008; Li et al., 2010; Jin et al., 2020b)— implicitly allude to specific exploration objectives, but do not appear to explicitly optimize them. We hope that by separating algorithms from objectives, our definition can bring clarity and better guide algorithm design going forward.

3. The L_1-Coverage Objective

This section introduces our main exploration objective, L_1-Coverage. Throughout the section, we work with a fixed (known) MDP M, and an arbitrary set of policies $\Pi \subseteq \Pi_{\mathsf{rns}}$.

L_1-Coverage objective. For a policy ensemble $p \in \Delta(\Pi_{\mathsf{rns}})$ and parameter $\varepsilon \in [0,1]$ we define the L_1-*Coverage* objective by

$$\Psi_{h,\varepsilon}^M(p) = \sup_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon\cdot d_h^{M,\pi}(x_h,a_h)}\right], \tag{2}$$

where we slightly overload notation for occupancy measures and write $d_h^{M,p}(x,a) := \mathbb{E}_{\pi\sim p}\big[d_h^{M,\pi}(x,a)\big]$.[2] This objective encourages the ensemble $p \in \Delta(\Pi)$ to cover the state space at least as well as any individual policy $\pi \in \Pi$, but in an *average-case* sense (with respect to the state distribution induced by π itself) that discounts hard-to-reach states. Importantly, L_1-Coverage only considers the *relative probability* of visiting states (that is, the ratio of occupancies), which is fundamentally different from "tabular" objectives such as Eq. (1) from prior work (Jin et al., 2020a; Hazan et al., 2019) and is essential to drive exploration in large state spaces. The approximation parameter $\varepsilon > 0$ allows one to discard regions of the state space that have low relative probability for all policies, removing the reachability assumption required by Eq. (1).

L_1-Coverage is closely related to previous *optimal design*-based objectives in online RL, and to several standard notions of coverage in offline RL. Indeed, L_1-Coverage can be viewed as a generalization of previously proposed optimal design-based objectives (Wagenmaker et al., 2022; Wagenmaker & Jamieson, 2022; Li et al., 2023; Mhammedi

[1] $J_R^M(\pi)$ denotes the expected reward of π in M under R.

[2] Likewise, $\mathbb{E}^{M,p}$ and $\mathbb{P}^{M,p}$ denote the expectation and law for the process where we sample policy $\pi \sim p$ and execute it in M.

et al., 2023a); see Appendices A and F for details. Regarding the latter, when $\varepsilon = 0$, $\Psi^M_{h,\varepsilon}(p)$ coincides with L_1-concentrability (Farahmand et al., 2010; Xie & Jiang, 2020; Zanette et al., 2021), and is equivalent to the χ^2-divergence between d^π and d^p up to a constant shift.

Before turning to algorithmic development, we first show that L_1-Coverage is indeed a valid exploration objective in the sense of Definition 2.1, then show that it satisfies desideratum (1), providing meaningful control over the intrinsic complexity of exploration.

L_1-Coverage enables downstream policy optimization. We prove that L_1-Coverage enables downstream policy optimization through a change-of-measure lemma, which shows that it is possible to transfer the expected value for any function g of interest (e.g., Bellman error) under any policy π to the expected value under p.

Proposition 3.1 (Change of measure for L_1-Coverage)**.** *For any distribution $p \in \Delta(\Pi_{\mathsf{rns}})$, we have that for all functions $g : \mathcal{X} \times \mathcal{A} \to [0, B]$, all $\pi \in \Pi$, and all $\varepsilon > 0$,*[3]

$$\mathbb{E}^{M,\pi}[g(x_h, a_h)] \leq 2\sqrt{\Psi^M_{h,\varepsilon}(p) \cdot \mathbb{E}^{M,p}[g^2(x_h, a_h)]} + \Psi^M_{h,\varepsilon}(p) \cdot (\varepsilon B). \tag{3}$$

Using this result, one can prove that, using data gathered from p, standard offline reinforcement learning algorithms such as Fitted Q-Iteration (`FQI`) succeed with sample complexity scaling with $\max_h \Psi^M_{h,\varepsilon}(p)$. One can similarly analyze hybrid offline/online methods (online methods that require access to exploratory data) such as `PSDP` (Bagnell et al., 2003) and `NPG` (Agarwal et al., 2021); see Appendix D for details. Such methods will prove to be critical for the reward-free reinforcement learning guarantees we present in the sequel.

In light of Proposition 3.1, which shows that L_1-Coverage satisfies Definition 2.1, we refer to any $p \in \Delta(\Pi_{\mathsf{rns}})$ that (approximately) optimizes the L_1-Coverage objective as a *policy cover* going forward.[4]

L_1-Coverability provides intrinsic complexity control. Of course, the guarantee in Proposition 3.1 is only useful if desideratum (1) is satisfied, i.e. there exist distributions $p \in \Delta(\Pi)$ for which the L_1-Coverage objective is bounded. To this end, we define the optimal value for the L_1-Coverage objective, which we refer to as L_1-*Coverability*, as

$$\mathsf{Cov}^M_{h,\varepsilon} = \inf_{p \in \Delta(\Pi)} \Psi^M_{h,\varepsilon}(p), \tag{4}$$

and define $\mathsf{Cov}^M_\varepsilon = \max_{h \in [H]} \mathsf{Cov}^M_{h,\varepsilon}$.

We show that the L_1-Coverability value $\mathsf{Cov}^M_{h,\varepsilon}$ can be interpreted as an intrinsic structural parameter for the MDP M, and is bounded for standard MDP classes of interest. To do so, we draw a connection a structural parameter introduced by Xie et al. (2023) as a means to bridge online and offline RL, which we refer to as L_∞-*Coverability*:

$$C^M_{\infty;h} = \inf_{\mu \in \Delta(\mathcal{X}\times\mathcal{A})} \sup_{\pi \in \Pi} \sup_{(x,a) \in \mathcal{X}\times\mathcal{A}} \left\{ \frac{d^{M,\pi}_h(x,a)}{\mu(x,a)} \right\}, \tag{5}$$

with $C^M_\infty = \max_{h \in [H]} C^M_{\infty;h}$.

L_∞-Coverability measures the best possible (worst-case) density ratio that can be achieved if one optimally designs the data distribution μ with knowledge of the underlying MDP. The main differences between the value $C^M_{\infty;h}$ and the L_1-Coverability value in Eq. (4) are that (i) Eq. (5) considers worst-case (L_∞-type) rather than average-case coverage, and (ii) Eq. (5) allows the distribution $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$ to be arbitrary, while Eq. (4) requires the distribution to be realized as a mixture of occupancies (in other words, Eq. (5) allows for *non-admissible* mixtures).[5] Due to the latter difference, it is unclear at first glance whether one can relate the two objectives, since allowing for non-admissible mixtures could potentially make the objective (5) much smaller. Nonetheless, the following result, which uses a non-trivial application of the minimax theorem, shows that L_1-Coverability is indeed bounded by C^M_∞.

Proposition 3.2. *For all $\varepsilon > 0$, we have $\mathsf{Cov}^M_{h,\varepsilon} \leq C^M_{\infty;h}$.*

Examples for which C^M_∞—and consequently $\mathsf{Cov}^M_\varepsilon$—is bounded by small problem-dependent constants include Block MDPs (Xie et al., 2023) ($C^M_\infty \leq |\mathcal{S}||\mathcal{A}|$, where $\mathcal{S}$ is the number of latent states), linear/low-rank MDPs (Huang et al., 2023) ($C^M_\infty \leq d|\mathcal{A}|$, where d is the feature dimension), and analytically sparse low-rank MDPs (Golowich et al., 2023) ($C^M_\infty \leq k|\mathcal{A}|$, where k is the sparsity level). Importantly, these examples (particularly Block MDPs and Low-Rank MDPs) require nonlinear function approximation. See Appendix H.2 for details; see also Xie et al. (2023); Amortila et al. (2024).

We note that L_1- and L_∞-Coverability are less general than structural parameters defined in terms of Bellman errors, such as Bellman rank (Jiang et al., 2017), Bellman-Eluder dimension (Jin et al., 2021a), and Bilinear rank (Du et al., 2021), which are known to not admit computationally efficient learning algorithms (Dann et al., 2018). In this sense, L_1-Coverage strikes a balance between generality and tractability. We discuss this in more detail and discuss connections to other structural parameters in Appendix F.

[3] This result is meaningful in the parameter regime where $\Psi^M_{h,\varepsilon}(p) < 1/\varepsilon$. We refer to this regime as *non-trivial*, as $\Psi^M_{h,\varepsilon}(p) \leq 1/\varepsilon$ holds vacuously for all p.

[4] This definition generalizes most notions of policy cover found in prior work (Du et al., 2019; Misra et al., 2020; Mhammedi et al., 2023b;a; Huang et al., 2023).

[5] We adopt the notation Cov^M for admissible variants of coverability and C^M for non-admissible variants of coverability.

Algorithm 1 Approximate Policy Cover Computation via L_∞-Coverability Relaxation

1: **input**: Layer $h \in [H]$, precision $\varepsilon \in [0,1]$, distribution $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$ w/ $C_\infty \equiv C^M_{\infty;h}(\mu)$, opt. tol. $\varepsilon_{\mathrm{opt}} > 0$.
2: Set $T = \frac{1}{\varepsilon}$.
3: **for** $t = 1, 2, \cdots, T$ **do**
4: Compute $\pi^t \in \Pi$ such that

$$\mathbb{E}^{M,\pi^t}\left[\frac{\mu(x_h,a_h)}{\sum_{i<t} d_h^{M,\pi^i}(x_h,a_h) + C_\infty \mu(x_h,a_h)}\right] \geq \sup_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{\sum_{i<t} d_h^{M,\pi^i}(x_h,a_h) + C_\infty \mu(x_h,a_h)}\right] - \varepsilon_{\mathrm{opt}}.$$

5: Return $p = \mathrm{Unif}(\pi^1, \ldots, \pi^T)$.

4. Optimizing L_1-Coverage: Efficient Planning

Directly optimizing the L_1-Coverage objective (2) presents challenges because the objective is quadratic in the occupancy $d^{M,\pi}$. To address this issue, this section provides two *relaxations*—that is, relaxed objectives that upper bound L_1-Coverage—that are directly amenable to optimization (via reduction to standard reward-driven policy optimization), yet are still bounded for MDPs of interest. Section 4.1 presents a relaxation based on a connection to L_∞-*Coverability* (Eq. (5)), and Section 4.2 presents a relaxation based on a connection to *Pushforward Coverability*. The first relaxation is tighter, but requires stronger knowledge of the underlying MDP. To motivate our results, recall that given a C-approximate minimizer $p \in \Delta(\Pi)$ for which $\Psi^M_{h,\varepsilon}(p) \leq C \cdot \mathsf{Cov}^M_{h,\varepsilon}$, the sample complexity in Proposition 3.1 degrades only by an $O(C)$ factor.

4.1. The L_∞-Coverability Relaxation

Our first relaxation of the L_1-Coverage objective assumes access to a distribution $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$ for which the L_∞-*concentrability coefficient* $C^M_{\infty;h}(\mu) := \sup_{\pi\in\Pi} \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \left\{\frac{d_h^{M,\pi}(x,a)}{\mu(x,a)}\right\}$ (Chen & Jiang, 2019) is bounded. For such a distribution μ, abbreviating $C_\infty \equiv C^M_{\infty;h}(\mu)$, we define

$$\Psi^M_{\mu;h,\varepsilon}(p) := \sup_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon \cdot C_\infty \mu(x_h,a_h)}\right], \tag{6}$$

and $\mathsf{Cov}^M_{\mu;h,\varepsilon} := \inf_{p\in\Delta(\Pi)} \Psi^M_{\mu;h,\varepsilon}(p)$. This objective upper bounds L_1-Coverage, and any $p \in \Delta(\Pi_{\mathsf{rns}})$ that optimizes it has $\Psi^M_{h,\varepsilon}(p) \leq 2C^M_{\infty;h}(\mu)$.

Proposition 4.1. *For a distribution μ with $C_\infty \equiv C^M_{\infty;h}(\mu)$, it holds that for all $p \in \Delta(\Pi_{\mathsf{rns}})$,*

$$\Psi^M_{h,\varepsilon}(p) \leq 2C_\infty \cdot \Psi^M_{\mu;h,\varepsilon}(p). \tag{7}$$

Furthermore, $\mathsf{Cov}^M_{\mu;h,\varepsilon} \leq 1$ for all $\varepsilon > 0$.

Notably, given access to a distribution $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$ that achieves the L_∞-Coverability value $C^M_{\infty;h}$ in Eq. (5), any distribution $p \in \Delta(\Pi_{\mathsf{rns}})$ that optimizes the relaxation in Eq. (6) achieves L_1-Coverage value $\Psi^M_{h,\varepsilon}(p) \leq 2C^M_{\infty;h}$. However, the relaxation supports arbitrary distributions $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$, allowing one to trade off approximation value and computation. Indeed, in some cases, it may be simpler to compute a distribution $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$ that has suboptimal, yet reasonable concentrability. Because μ is not required to be admissible, such a distribution can be computed easily or in closed form for many MDP families of interest. For example, in tabular MDPs we can simply take $\mu = \mathrm{Unif}(\mathcal{X} \times \mathcal{A})$. See Appendix H.2 for more examples.

The algorithm. Algorithm 1 provides an iterative algorithm to compute a distribution $p \in \Delta(\Pi_{\mathsf{rns}})$ that optimizes the L_∞-relaxation in Eq. (6). The algorithm proceeds in T steps. At each step $t \in [T]$, given a sequence of policies $\pi^1, \ldots, \pi^{t-1}$ computed so far, the algorithm computes a new policy π^t by solving the policy optimization problem

$$\pi^t = \arg\max_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{\sum_{i<t} d_h^{M,\pi^i}(x_h,a_h) + C_\infty \mu(x_h,a_h)}\right] \tag{8}$$

in Line 4 (up to tolerance $\varepsilon_{\mathrm{opt}} > 0$). After all T rounds conclude, the algorithm returns the uniform mixture $p = \mathrm{Unif}(\pi^1, \ldots, \pi^T)$ as a policy cover.

The optimization problem (8) aims to find a policy π^t that explores regions of the state space not already covered by $\pi^1, \ldots, \pi^{t-1}$. Critically, it is a standard *reward-driven policy optimization* problem with reward function

$$r^t_h(x,a) := \frac{\mu(x,a)}{\sum_{i<t} d_h^{M,\pi^i}(x,a) + C_\infty \mu(x,a)}.$$

In practice, one can solve Eq. (8) using standard policy optimization algorithms (we take this approach in Section 6). However, since the MDP M is known, we can also take advantage of the vast literature on algorithms for planning with a known model, as well as algorithms like Policy Search by Dynamic Programming (Bagnell et al. (2003)) or Natural Policy Gradient (Agarwal et al. (2021)) that can use μ itself as a high-quality reset distribution.

The following theorem shows that Algorithm 1 converges to a policy cover $p \in \Delta(\Pi_{\mathsf{rns}})$ that optimizes the relaxation in Eq. (6) (up to a small $\log(\varepsilon^{-1})$ multiplicative approximation factor) in a small number of iterations.

Theorem 4.1. *For any $\varepsilon \in [0,1]$ and $h \in [H]$, given a distribution μ with $C_\infty \equiv C^M_{\infty;h}(\mu)$, whenever $\varepsilon_{\mathrm{opt}} \leq$*

$\varepsilon \log(2\varepsilon^{-1})$, *Algorithm 1 with* $T = \varepsilon^{-1}$ *produces a distribution* $p \in \Delta(\Pi)$ *with* $|\mathrm{supp}(p)| \le \varepsilon^{-1}$ *such that*

$$\Psi^M_{\mu;h,\varepsilon}(p) \le 3\log(2\varepsilon^{-1}), \tag{9}$$

and consequently $\Psi^M_{h,\varepsilon}(p) \le 6C_\infty \log(2\varepsilon^{-1})$.

See Appendix H.2 for examples. Note that the algorithmic template in Algorithm 1 can be applied in unknown MDPs as long as Line 4 can be implemented sample-efficiently; this observation is the basis for the model-free online exploration algorithm given in Appendix E.

4.2. The Pushforward Coverability Relaxation

The main drawback behind the L_∞-Coverability relaxation in the prequel is the assumption of access to a covering distribution $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$.[6] The next relaxation, which is inspired by the notion of *pushforward concentrability* in offline reinforcement learning (Xie & Jiang, 2021; Foster et al., 2022), removes this assumption at the cost of giving a looser upper bound. This objective takes the form

$$\Psi^M_{\mathsf{push};h,\varepsilon}(p) = \sup_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{d^{M,p}_h(x_h) + \varepsilon \cdot P^M_{h-1}(x_h \mid x_{h-1}a_{h-1})}\right], \tag{10}$$

with $\mathsf{Cov}^M_{\mathsf{push};h,\varepsilon} := \inf_{p\in\Delta(\Pi)} \Psi^M_{\mathsf{push};h,\varepsilon}(p)$. This objective replaces the covering distribution μ with the transition distribution P^M_{h-1} for M, making it more practical to compute. Unlike the L_∞-Coverability relaxation (6), the optimal value $\mathsf{Cov}^M_{\mathsf{push};h,\varepsilon}$ may not be bounded by the parameter $C^M_{\infty;h}$. However, we show that the value can be controlled by a related *pushforward coverability* parameter given by $C^M_{\mathsf{push};h} = \inf_{\mu\in\Delta(\mathcal{X})} \sup_{(x,a,x')\in\mathcal{X}\times\mathcal{A}\times\mathcal{X}} \left\{\frac{P^M_{h-1}(x'|x,a)}{\mu(x')}\right\}$, with $C^M_{\mathsf{push}} = \max_{h\in[H]} C^M_{\mathsf{push};h}$.[7]

Proposition 4.2. *Fix* $h \in [H]$. *For any* $p \in \Delta(\Pi_{\mathsf{rns}})$, *if we define* $p' \in \Delta(\Pi_{\mathsf{rns}})$ *as the distribution induced by sampling* $\pi \sim p$ *and executing* $\pi \circ_h \pi_{\mathrm{unif}}$, *we have that for all* $\varepsilon > 0$,

$$\Psi^M_{h,\varepsilon}(p') \le |\mathcal{A}| \cdot \Psi^M_{\mathsf{push};h,\varepsilon}(p). \tag{11}$$

Furthermore, $\mathsf{Cov}^M_{\mathsf{push};h,\varepsilon} \le C^M_{\mathsf{push};h}$ *for all* $\varepsilon > 0$.

In particular, any distribution $p \in \Delta(\Pi_{\mathsf{rns}})$ that optimizes the relaxation in Eq. (10) achieves L_1-Coverage value $\Psi^M_{h,\varepsilon}(p') \le |\mathcal{A}| \cdot C^M_{\mathsf{push};h}$.[8] Notable special cases where pushforward coverability is bounded include: (i) Tabular MDPs have $C^M_{\mathsf{push}} \le |\mathcal{X}|$, (ii) Block MDPs (Du et al., 2019; Misra et al., 2020; Zhang et al., 2022; Mhammedi et al., 2023b) with latent state space $\mathcal{S}$ have $C^M_{\mathsf{push}} \le |\mathcal{S}|$, (iii) Low-Rank MDPs (30) have $C^M_{\mathsf{push}} \le d$ (Xie & Jiang, 2021).

[6] Note that since we consider planning, this is a computational assumption, not a statistical assumption.

[7] This is inspired by *pushforward concentrability* (Xie & Jiang, 2021); $C^M_{\infty;h} \le |\mathcal{A}| \cdot C^M_{\mathsf{push};h}$ but the converse is not true.

[8] The dependence on $|\mathcal{A}|$ in this result is natural, as pushforward coverability only considers state occupancies.

The algorithm. An iterative algorithm to optimize the pushforward coverability relaxation is given in Algorithm 5 (deferred to Appendix H.1 for space). The algorithm follows the same template as Algorithm 1, but at each step t, the policy π^t is computed by solving an objective based on Eq. (10). The main guarantee is as follows.

Theorem 4.2. *For any* $\varepsilon \in [0,1]$ *and* $h \in [H]$, *whenever* $\varepsilon_{\mathrm{opt}} \le C^M_{\mathsf{push};h} \cdot \varepsilon \log(2\varepsilon^{-1})$, *Algorithm 5 produces a distribution* $p \in \Delta(\Pi)$ *with* $|\mathrm{supp}(p)| \le \varepsilon^{-1}$ *such that*

$$\Psi^M_{\mathsf{push};h,\varepsilon}(p) \le 5C^M_{\mathsf{push};h} \log(2\varepsilon^{-1}). \tag{12}$$

Consequently, if we define $p' \in \Delta(\Pi_{\mathsf{rns}})$ *as the distribution induced by sampling* $\pi \sim p$ *and executing* $\pi \circ_h \pi_{\mathrm{unif}}$, *we have that* $\Psi^M_{h,\varepsilon}(p') \le 5|\mathcal{A}|C^M_{\mathsf{push};h} \log(2\varepsilon^{-1})$.

5. Efficient Online Exploration via L_1-Coverage

In this section, we turn our attention to sample-efficient *online exploration* for the setting in which the underlying MDP $M^\star$ is unknown. Throughout the section, we work with an arbitrary user-specified subset of policies $\Pi \subseteq \Pi_{\mathsf{rns}}$.

Model-based reinforcement learning setup. We focus on *model-based reinforcement learning*, and assume access to a *model class* $\mathcal{M}$ that contains the true MDP $M^\star$.

Assumption 5.1 (Realizability)**.** *The learner has access to a class* $\mathcal{M}$ *containing the true model* $M^\star$.

The class $\mathcal{M}$ is user-specified, and can be parameterized by deep neural networks or any other flexible function class, with the best choice depending on the problem domain.

For $M \in \mathcal{M}$, we use $M(\pi)$ as shorthand for the law of the trajectory $(x_1, a_1), \ldots, (x_H, a_H)$ for policy π in M. We define the squared Hellinger distance for probability measures $\mathbb{P}$ and $\mathbb{Q}$ by $D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) = \int \left(\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}}\right)^2$.

Estimation oracles. Our algorithms and regret bounds use the primitive of an *estimation oracle*, denoted by $\mathbf{Alg}_{\mathsf{Est}}$, a user-specified algorithm for estimation that is used to estimate the underlying model $M^\star$ from data (Foster & Rakhlin, 2020; 2023) sequentially. At each episode t, given the data $\mathfrak{H}^{t-1} = (\pi^1, o^1), \ldots, (\pi^{t-1}, o^{t-1})$ collected so far, where $o^t := (x^t_1, a^t_1), \ldots, (x^t_H, a^t_H)$, the estimation oracle constructs an estimate $\widehat{M}^t = \mathbf{Alg}_{\mathsf{Est}}\left(\{(\pi^i, o^i)\}_{i=1}^{t-1}\right)$ for the true MDP $M^\star$. We assume that $\mathbf{Alg}_{\mathsf{Est}}$ is an *offline* estimation oracle, in the sense that each estimator $\widehat{M}^t$ has good out-of-sample performance on the historical dataset $\mathfrak{H}^{t-1}$.

Assumption 5.2 (Offline estimation oracle)**.** *At each time* $t \in [T]$, *an offline estimation oracle* $\mathbf{Alg}_{\mathsf{Est}}$ *for* $\mathcal{M}$, *given* $\mathfrak{H}^{t-1} = (\pi^1, o^1), \ldots, (\pi^{t-1}, o^{t-1})$ *with* $o^i \sim M^\star(\pi^i)$ *and* $\pi^i \sim p^i$, *returns an estimator* $\widehat{M}^t \in \mathcal{M}$ *such that*

Algorithm 2 Coverage-Driven Exploration (CODEX)

1: **input**: Estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$, number of episodes $T \in \mathbb{N}$, approximation parameters $C \geq 1$, $\varepsilon \in [0,1]$.
2: **for** $t = 1, 2, \cdots, T$ **do**
3: Estimate model: $\widehat{M}^t = \mathbf{Alg}^t_{\mathsf{Est}}\big(\{(\pi^i, o^i)\}_{i=1}^{t-1}\big)$.
// Plug-in approximation to L_1-Coverage objective.
4: For each $h \in [H]$, compute (C,ε)-approx. policy cover p^t_h for $\widehat{M}^t$: $\Psi^{\widehat{M}^t}_{h,\varepsilon}(p^t_h) =$

$$\sup_{\pi\in\Pi} \mathbb{E}^{\widehat{M}^t,\pi}\left[\frac{d^{\widehat{M}^t,\pi}_h(x_h,a_h)}{d^{\widehat{M}^t,p^t_h}_h(x_h,a_h) + \varepsilon \cdot d^{\widehat{M}^t,\pi}_h(x_h,a_h)}\right] \leq C. \tag{13}$$

5: Sample $\pi^t \sim q^t = \mathrm{Unif}(p^t_1, \ldots, p^t_H)$ and observe trajectory $o^t = (x^t_1, a^t_1), \ldots, (x^t_H, a^t_H)$.
6: **return** $(p_1, \ldots, p_H)$, where $p_h := \mathrm{Unif}(p^1_h, \ldots, p^T_h)$.

$\mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(t) :=$

$$\sum_{i<t} \mathbb{E}_{\pi^i\sim p^i}\big[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi^i), M^\star(\pi^i)\big)\big] \leq \mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(\mathcal{M}, T, \delta),$$

for all $t \in [T]$ with probability at least $1-\delta$ whenever $M^\star \in \mathcal{M}$, where $\mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(\mathcal{M}, T, \delta)$ is a known upper bound.

As an example, the standard maximum likelihood estimator (MLE) satisfies Assumption 5.2 with $\mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(\mathcal{M}, T, \delta) \leq O(\log(|\mathcal{M}|T/\delta))$ (e.g., Foster & Rakhlin (2023)).

Algorithm. Our main algorithm for reward-free reinforcement learning, Coverage-Driven Exploration (CODEX; Algorithm 2), is based on a simple "plug-in" estimation-optimization paradigm: Repeatedly compute an estimate $\widehat{M}^t$ for the true model $M^\star$, then compute a policy cover $p \in \Delta(\Pi)$ that optimizes the L_1-Coverage objective for $\widehat{M}^t$ (a *plug-in* approximation to the true L_1-Coverage objective) and use this to collect data; proceed until this process arrives at a high-quality cover for $M^\star$.

In more detail, Algorithm 2 proceeds in T episodes. At each episode t, the algorithm invokes the user-specified estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$ to produce an estimate $\widehat{M}^t$ for the model $M^\star$ based on the data collected so far. Given this estimate, for each layer $h \in [H]$, the algorithm computes a (C,ε)-approximate policy cover $p^t_h \in \Delta(\Pi)$ for $\widehat{M}^t$: $\Psi^{\widehat{M}^t}_{h,\varepsilon}(p^t_h) =$

$$\sup_{\pi\in\Pi} \mathbb{E}^{\widehat{M}^t,\pi}\left[\frac{d^{\widehat{M}^t,\pi}_h(x_h,a_h)}{d^{\widehat{M}^t,p^t_h}_h(x_h,a_h) + \varepsilon \cdot d^{\widehat{M}^t,\pi}_h(x_h,a_h)}\right] \leq C. \tag{14}$$

where $C > 0$ is made sufficiently large to ensure Eq. (14) is feasible. This is a plug-in approximation to the true L_1-Coverage objective $\Psi^{M^\star}_{h,\varepsilon}(p)$. Given the approximate covers $p^t_1, \ldots, p^t_H$, the algorithm collects a new trajectory o^t by sampling $\pi^t \sim q^t := \mathrm{Unif}(p^t_1, \ldots, p^t_H)$. This trajectory is used to update the estimator $\widehat{M}^t$, and the algorithm proceeds to the next episode. Once all episodes conclude, the algorithm returns $p_h := \mathrm{Unif}(p^1_h, \ldots, p^T_h)$ as the final cover for each layer h.

The plug-in L_1-Coverage objective in Eq. (14) can be solved efficiently using the relaxation-based methods in Section 4, since the model $\widehat{M}^t \in \mathcal{M}$ is known, making this a pure (non-statistical) planning problem. We leave the approximation parameter $C \geq 1$ as a free parameter to accommodate the approximation factors these relaxations incur (the sample complexity degrades linearly with C).

Main result. In its most general form, Algorithm 2 leads to sample complexity guarantees based on L_1-Coverability. Due to space constraints, the main result we present here is a simpler guarantee based on L_∞-Coverability. To state the guarantee in the most compact form possible, we make the following assumption on the estimation error rate.

Assumption 5.3 (Parametric estimation rate)**.** *The offline estimation oracle satisfies $\mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(\mathcal{M}, T, \delta) \leq O(d_{\mathsf{est}} \log(B_{\mathsf{est}} T/\delta))$ for parameters $d_{\mathsf{est}}, B_{\mathsf{est}} \in \mathbb{N}$.*

Theorem 5.1 (Main guarantee for CODEX)**.** *Let $\varepsilon > 0$ be given. Let $C_\infty \equiv C^{M^\star}_\infty$, and suppose that (i) we restrict $\mathcal{M}$ such that all $M \in \mathcal{M}$ have $C^M_\infty \leq C_\infty$, and (ii) we solve Eq. (13) with $C = C_\infty$ for all t.*[9] *Then, given an offline estimation oracle satisfying Assumptions 5.2 and 5.3, using $T = \widetilde{O}\Big(\frac{H^8 (C^{M^\star}_\infty)^3 d_{\mathsf{est}} \log(B_{\mathsf{est}}/\delta)}{\varepsilon^4}\Big)$ episodes, Algorithm 2 produces policy covers $p_1, \ldots, p_H \in \Delta(\Pi)$ such that*

$$\forall h \in [H]: \quad \Psi^{M^\star}_{h,\varepsilon}(p_h) \leq 12H \cdot C^{M^\star}_\infty \tag{15}$$

w.p. at least $1-\delta$. For a finite class $\mathcal{M}$, if we use MLE as the estimator, we can take $T = \widetilde{O}\Big(\frac{H^8 (C^{M^\star}_\infty)^3 \log(|\mathcal{M}|/\delta)}{\varepsilon^4}\Big)$.

Theorem 5.1 shows for the first time that it is possible to perform sample-efficient and computationally efficient reward-free exploration under coverability. We refer to Appendix I.4 for the proof, which is based on a connection to the Decision-Estimation Coefficient of Foster et al. (2021; 2023). Some key features of the result are as follows.

Computational efficiency. As maximum likelihood estimation (MLE) is a valid estimation oracle, Algorithm 2 is computationally efficient whenever 1) MLE can be performed efficiently, and 2) the plug-in L_1-Coverage objective in Eq. (13) can be approximately optimized efficiently. As the objective involves only the estimated model $\widehat{M}^t$ (and hence is a computational problem), we can use the relaxations in Section 4 to do this efficiently (via off-the-shelf methods).

Regarding the latter point, note that while Theorem 5.1 assumes for simplicity that Eq. (13) is solved with $C = C^{M^\star}_\infty$,

[9] We can take $C^M_\infty \leq C_\infty$ w.l.o.g. when C_∞ is known. In this case, solving Eq. (13) with $C = C_\infty$ is feasible by Proposition 3.2.

if we solve the objective for $C > C_\infty^{M^\star}$ the result continues to hold with $C_\infty^{M^\star}$ replaced by C in the sample complexity bound and approximation guarantee. Consequently, if we solve Eq. (13) using Algorithm 1, the guarantees in Theorem 5.1 continue to hold up to an $\widetilde{O}(1)$ approximation factor.

Statistical efficiency. CODEX achieves sample complexity guarantees based on L_∞-Coverability in a computationally eficient fashion for the first time. Compared to previous inefficient algorithms based on coverability (Xie et al., 2023; Liu et al., 2023a; Amortila et al., 2024) our theorem has somewhat looser sample complexity, and requires model-based function approximation. See Appendix E for a model-free counterpart based on weight function learning.

CODEX is not limited to L_∞-Coverability. Appendix I.1 gives more general results (Theorem I.1, Corollary I.1) which achieve sample complexity $\mathrm{poly}(\mathsf{Cov}_\varepsilon^{M^\star}, H, \log|\mathcal{M}|, \varepsilon^{-1})$, showing that L_1-Coverability is itself a sufficiently powerful structural parameter to enable sample-efficient learning with nonlinear function approximation.

Application to downstream policy optimization. By Proposition 3.1, the policy covers $p_{1:H}$ returned by Algorithm 2 can be used to optimize any downstream reward function using offline RL, giving end-to-end guarantees for reward-driven online RL. We sketch an example using MLE for policy optimization in Appendix I.2 (cf. Appendix D).

6. Experimental Evaluation

We present proof-of-concept experiments to validate our theoretical results.[10] We focus on the *planning* problem (Section 4), and consider the classical MountainCar environment (Brockman et al., 2016). We optimize the L_1-Coverage objective via Algorithm 1 and compare this to the MaxEnt algorithm (Hazan et al., 2019) and uniform exploration. We find that L_1-Coverage explores the state space more quickly and effectively than the baselines.

Experimental setup. MountainCar is a continuous domain with two-dimensional states and actions. We consider a deterministic starting state near the bottom of the "valley" in the environment, so that deliberate exploration is required.

We optimize L_1-Coverage using Algorithm 1, approximating the occupancies d^{π^i} via count-based estimates on a discretization of the state-action space. We set μ to the uniform distribution and C_∞ to the number of state-action pairs (in the discretization). For MaxEnt, we use the implementation from Hazan et al. (2019). Algorithm 1 and MaxEnt both require access to a subroutine for reward-driven planning (to solve the problem in Eq. (8) or an analogous subproblem). For both, we use a policy gradient method (REINFORCE; Sutton et al. (1999)) as the reward-driven planner, following Hazan et al. (2019); our policy class is parameterized as a two layer neural network. For details on the environment, architectures, and hyperparameters, see Appendix B.

Results. Results are visualized in Figure 1. We measure the number of unique discretized states discovered by each algorithm's policy covers, and find that L_1-Coverage (Algorithm 1) outperforms the MaxEnt and uniform exploration baselines by a factor of two. We also perform a qualitative comparison by visualizing the occupancies induced by the learned policy covers, and find that the cover obtained with L_1-Coverage explores a much larger portion of the state space than the baselines; notably, L_1-Coverage explores both hills in the MountainCar environment, while the baselines fail to do so. We find these results promising, and plan to perform a large-scale evaluation in future work; see Appendix B for further experimental results.

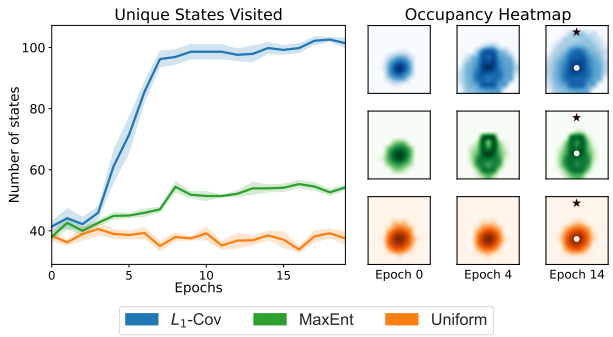

Figure 1. Unique discrete states visited (mean/standard error over 10 runs) and occupancy heatmaps for each policy cover obtained by L_1-Coverage (Algorithm 1), MaxEnt, and uniform actions. Each epoch comprises one policy cover update. Heatmap legend: velocity (x-axis), position (y-axis), start state (•), goal state (⋆).

7. Discussion and Open Problems

Our results show that the L_1-Coverage objective serves as a scalable primitive for exploration in reinforcement learning, providing sample efficiency, computational efficiency, and flexibility. On the theoretical side, our results raise a number of interesting questions for future work: (i) What are the weakest representation conditions under which coverability leads to computationally efficient online reinforcement learning guarantees? (ii) Can we generalize the L_1-Coverage objective further while still allowing for practical/computationally efficient optimization?

Additional results. Several additional results are deferred to the appendix due to space constraints. Appendix E builds on Sections 4 and 5 to provide efficient algorithms for reward-free exploration under a weaker form of *model-free* function approximation based on density ratio realizability, and Appendix F provides connections between L_1-Coverability and other intrinsic structural properties.

[10] Code available at github.com/philip-amortila/l1-coverability.

Impact Statement

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

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, 2014.

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

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 2021.

Amortila, P., Foster, D. J., Jiang, N., Sekhari, A., and Xie, T. Harnessing density ratios for online reinforcement learning. *International Conference on Learning Representations*, 2024.

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

Ash, J. T., Zhang, C., Goel, S., Krishnamurthy, A., and Kakade, S. Anti-concentrated confidence bonuses for scalable exploration. In *International Conference on Learning Representations*, 2022.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 2002.

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

Bagnell, J., Kakade, S. M., Schneider, J., and Ng, A. Policy search by dynamic programming. *Advances in Neural Information Processing Systems*, 2003.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. *Advances in Neural Information Processing Systems*, 2016.

Block, A. and Polyanskiy, Y. The sample complexity of approximate rejection sampling with applications to smoothed online learning. In *Conference on Learning Theory*, 2023.

Block, A., Dagan, Y., Golowich, N., and Rakhlin, A. Smoothed online learning is as easy as statistical learning. In *Conference on Learning Theory*, 2022.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI gym. *arXiv:1606.01540*, 2016.

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation. In *International Conference on Learning Representations*, 2018.

Chen, F., Mei, S., and Bai, Y. Unified algorithms for RL with decision-estimation coefficients: No-regret, PAC, and reward-free learning. *arXiv:2209.11745*, 2022a.

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

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

Chen, J., Modi, A., Krishnamurthy, A., Jiang, N., and Agarwal, A. On the statistical efficiency of reward-free exploration in non-linear rl. *Advances in Neural Information Processing Systems*, 2022b.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. On oracle-efficient PAC RL with rich observations. In *Advances in Neural Information Processing Systems*, 2018.

Daskalakis, C., Golowich, N., Haghtalab, N., and Shetty, A. Smooth nash equilibria: Algorithms and complexity. In *Innovations in Theoretical Computer Science*, 2023.

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

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

Farahmand, A.-m., Szepesvári, C., and Munos, R. Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 2010.

Foster, D. J. and Krishnamurthy, A. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Neural Information Processing Systems*, 2021.

Foster, D. J. and Rakhlin, A. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *International Conference on Machine Learning*, 2020.

Foster, D. J. and Rakhlin, A. Foundations of reinforcement learning and interactive decision making. *arXiv:2312.16730*, 2023.

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

Foster, D. J., Krishnamurthy, A., Simchi-Levi, D., and Xu, Y. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory*, 2022.

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

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2010.

Golowich, N., Rohatgi, D., and Moitra, A. Exploring and learning in sparse linear mdps without computationally intractable oracles. *arXiv:2309.09457*, 2023.

Haghtalab, N., Roughgarden, T., and Shetty, A. Smoothed analysis of online and differentially private learning. *Advances in Neural Information Processing Systems*, 2020.

Haghtalab, N., Han, Y., Shetty, A., and Yang, K. Oracle-efficient online learning for smoothed adversaries. *Advances in Neural Information Processing Systems*, 2022a.

Haghtalab, N., Roughgarden, T., and Shetty, A. Smoothed analysis with adaptive adversaries. In *Symposium on Foundations of Computer Science (FOCS)*, 2022b.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, 2019.

Huang, A., Chen, J., and Jiang, N. Reinforcement learning in low-rank MDPs with density features. *International Conference on Machine Learning*, 2023.

Jiang, N. and Huang, J. Minimax value interval for off-policy evaluation and policy optimization. *Neural Information Processing Systems*, 2020.

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

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? *Advances in Neural Information Processing Systems*, 2018.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, 2020a.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2020b.

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

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, 2021b.

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

Katdare, P., Jiang, N., and Driggs-Campbell, K. Marginalized importance sampling for off-environment policy evaluation. In *Conference on Robot Learning*, 2023.

Kiefer, J. and Wolfowitz, J. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 1960.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Kober, J., Bagnell, J. A., and Peters, J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 2013.

Li, G., Zhan, W., Lee, J. D., Chi, Y., and Chen, Y. Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., and Gao, J. Deep reinforcement learning for dialogue generation. In *Conference on Empirical Methods in Natural Language Processing*, 2016.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web*, 2010.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2015.

Liu, F., Viano, L., and Cevher, V. Provable benefits of general coverage conditions in efficient online RL with function approximation. *International Conference on Machine Learning*, 2023a.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 2018.

Liu, Z., Lu, M., Xiong, W., Zhong, H., Hu, H., Zhang, S., Zheng, S., Yang, Z., and Wang, Z. Maximize to explore: One objective function fusing estimation, planning, and exploration. In *Advances in Neural Information Processing Systems*, 2023b.

Martin, J., Sasikumar, S. N., Everitt, T., and Hutter, M. Count-based exploration in feature space for reinforcement learning. In *International Joint Conference on Artificial Intelligence*, 2017.

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

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

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

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 2015.

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

Munos, R. Error bounds for approximate policy iteration. In *International Conference on Machine Learning*, 2003.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 2010.

Ozdaglar, A. E., Pattathil, S., Zhang, J., and Zhang, K. Revisiting the linear-programming framework for offline RL with general function approximation. In *International Conference on Machine Learning*, 2023.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 2019.

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

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 2021.

Rashidinejad, P., Zhu, H., Yang, K., Russell, S., and Jiao, J. Optimal conservative offline RL with general function approximation via augmented lagrangian. In *International Conference on Learning Representations*, 2023.

Rendle, S., Freudenthaler, C., and Schmidt-Thieme, L. Factorizing personalized Markov chains for next-basket recommendation. In *International Conference on World Wide Web*, 2010.

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

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 1958.

Song, Y., Zhou, Y., Sekhari, A., Bagnell, D., Krishnamurthy, A., and Sun, W. Hybrid RL: Using both offline and online data can make RL efficient. In *International Conference on Learning Representations*, 2022.

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

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 1999.

Tang, H., Houthooft, R., Foote, D., Stooke, A., Xi Chen, O., Duan, Y., Schulman, J., DeTurck, F., and Abbeel, P. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 2017.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, 2020.

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W., and Xie, T. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv:2102.02981*, 2021.

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

Wagenmaker, A. and Jamieson, K. G. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *Advances in Neural Information Processing Systems*, 2022.

Wagenmaker, A. and Pacchiano, A. Leveraging offline data in online reinforcement learning. In *International Conference on Machine Learning*, 2023.

Wagenmaker, A. J., Simchowitz, M., and Jamieson, K. Beyond no regret: Instance-dependent PAC reinforcement learning. In *Conference on Learning Theory*, 2022.

Wang, R., Du, S. S., Yang, L., and Salakhutdinov, R. R. On reward-free reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 2020a.

Wang, R., Salakhutdinov, R. R., and Yang, L. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 2020b.

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

Xie, T. and Jiang, N. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, 2021.

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

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 2020.

Yao, H., Szepesvári, C., Pires, B. Á., and Zhang, X. Pseudo-MDPs and factored linear action models. In *Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, 2014.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 2021.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, 2022.

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

Contents of Appendix

A Additional Related Work **14**

B Experimental Evaluation: Details and Additional Results **15**
B.1 Experimental Details . . . 15
B.2 Additional Experimental Results . . . 16
B.3 Additional Discussion . . . 19

C Technical Tools **20**
C.1 Minimax Theorem . . . 20
C.2 Concentration . . . 20
C.3 Information Theory . . . 20
C.4 Reinforcement Learning . . . 20
C.5 Helper Lemmas . . . 21

I Additional Results and Discussion **22**

D L_1-Coverage: Application to Downstream Policy Optimization **22**
D.1 Proofs . . . 23

E Efficient Model-Free Exploration via L_1-Coverage **25**
E.1 Algorithm . . . 25
E.2 Main Result . . . 28

F L_1-Coverage: Structural Properties **29**
F.1 Connection to Non-Admissible L_1-Coverage . . . 29
F.2 Connection to Feature Coverage . . . 29
F.3 Insufficiency of Alternative Notions of Coverage . . . 30

II Proofs **31**

G Proofs from Section 3 **31**

H Proofs and Additional Details from Section 4 **32**
H.1 Omitted Algorithms . . . 32
H.2 Examples for Algorithm 1 . . . 32
H.3 Proofs from Section 4.1 . . . 33
H.4 Proofs from Section 4.2 . . . 34

I Proofs and Additional Details from Section 5 **36**
I.1 General Guarantees for Algorithm 2 . . . 37
I.2 Applying Algorithm 2 to Downstream Policy Optimization . . . 38
I.3 Technical Lemmas . . . 38
I.4 Proofs from Section 5 and Appendix I.1 . . . 41

J Proofs from Appendix F **47**

K Proofs from Appendix E **49**
K.1 General Guarantees for Algorithm 3 . . . 49
K.2 Technical Preliminaries . . . 51
K.3 Weight Function Estimation . . . 52
K.4 Policy Optimization Subroutines . . . 54
K.5 Proof of Theorem K.1 . . . 55

A. Additional Related Work

This section discusses additional related work not already covered in detail.

Theoretical objectives for exploration. On the computational side, our work can be viewed as building on a recent line of research that tries to make exploration computationally efficient by inductively building policy covers and using them to guide exploration. Up to this point, Block MDPs (Du et al., 2019; Misra et al., 2020; Zhang et al., 2022; Mhammedi et al., 2023b) and Low-Rank MDPs (Agarwal et al., 2020; Modi et al., 2024; Uehara et al., 2022; Zhang et al., 2022; Mhammedi et al., 2023b) are the most general classes that have been addressed by this line of research. Our work expands the scope of problems for which efficient exploration is possible beyond these classes to include the more general coverability framework. This is meaningful generalization, as it allows for nonlinear transition dynamics for the first time.

The L_1-Coverage objective can be viewed as a generalization of *optimal design*, an exploration objective which has previously been studied in the context of tabular and Low-Rank MDPs (Wagenmaker et al., 2022; Wagenmaker & Jamieson, 2022; Li et al., 2023; Mhammedi et al., 2023a); Appendix F.2 makes this connection explicit. In particular, Li et al. (2023) consider an optimal design objective for tabular MDPs which is equivalent to Eq. (6) with $\mu = \mathrm{Unif}(\mathcal{X})$. Outside of this paper, the only work we are aware of that considers optimal design-like objectives for nonlinear settings beyond Low-Rank MDPs is Foster et al. (2021), but their algorithms are not efficient in terms of offline estimation oracles.

Lastly, a theoretical exploration objective worth mentioning is maximum entropy exploration (Hazan et al., 2019; Jin et al., 2020a). Existing theoretical guarantees for this objective are limited to tabular MDPs, and we suspect the objective is not sufficiently powerful to allow for downstream policy optimization in more general classes of MDPs.

Coverability. Compared to previous guarantees based on coverability (Xie et al., 2023; Liu et al., 2023a; Amortila et al., 2024) our guarantees have somewhat looser sample complexity and require stronger function approximation. However, these works only consider reward-driven exploration and are computationally inefficient. A second contribution of our work is to show that L_1-Coverability, which can be significantly weaker than L_∞-Coverability, is sufficient for sample-efficient online reinforcement learning on its own.[11]

Instance-dependent algorithms and complexity. Wagenmaker et al. (2022); Wagenmaker & Jamieson (2022) provide instance-dependent regret bounds for tabular MDPs and linear MDPs that scale with problem-dependent quantities closely related to L_1-Coverability. These results are tailored to the linear setting, and their bounds contain lower-order terms that scale explicitly with the dimension and/or number of states.

General-purpose complexity measures. A long line of research provides structural complexity measures that enable sample-efficient exploration in reinforcement learning (Russo & Van Roy, 2013; Jiang et al., 2017; Sun et al., 2019; Wang et al., 2020b; Du et al., 2021; Jin et al., 2021a; Foster et al., 2021; Xie et al., 2023; Foster et al., 2023). Coverability is incomparable to most prior complexity measures (Xie et al., 2023), but is subsumed by the Decision-Estimation Coefficient (Foster et al., 2021), as well as the (less general) Sequential Extrapolation Coefficient and related complexity measures (Liu et al., 2023b). The main contrast between these works and our own is that they do not provide computationally efficient algorithms in general.

A handful of works extend the approaches above to accommodate reward-free reinforcement learning, but are still computationally inefficient, and do not explicitly suggest exploration objectives (Chen et al., 2022b; Xie et al., 2023; Chen et al., 2022a).

Further related work. Coverability is closely related to a notion of *smoothness* used in a line of recent work on smoothed online learning and related problems (Haghtalab et al., 2020; 2022a;b; Block et al., 2022; Block & Polyanskiy, 2023; Daskalakis et al., 2023). We are not aware of explicit technical connections between our techniques and these works, but it would be interesting to explore this in more detail.

[11] Liu et al. (2023a) also provide guarantees based on L_1-Coverability, but their regret bounds contain a lower-order term that can be as large as the number of states in general.

B. Experimental Evaluation: Details and Additional Results

B.1. Experimental Details

`MaxEnt` baseline. We implement our algorithm by building on top of the codebase for `MaxEnt` from Hazan et al. (2019). The `MaxEnt` algorithm follows a similar template to ours, in that it iteratively defines a reward function based on past occupancies and optimizes it to find a new exploratory policy. In `MaxEnt`, the reward function is defined as

$$r_t = \frac{dD_{\mathsf{KL}}(\mathrm{Unif} \,\|\, X)}{dX}\Big|_{X=\widehat{d}^{p^t}},$$

where $\widehat{d}^{p^t}$ is the estimated occupancy for the policy cover p^t at time t and Unif denotes the uniform distribution over $\mathcal{X} \times \mathcal{A}$. We use the same neural network architecture (described in detail below) to represent policies, the same algorithm for approximate planning, the same discretization scheme to approximate occupancies, and the exact same hyperparameters for discretization resolution, number of training steps, length of rollouts, learning rates, and so on. We found that our method worked "out-of-the-box" without any modifications to their architecture or hyperparameters.

Starting from their codebase, we obtain an implementation of Algorithm 1 simply by modifying the reward function given to the planner from the `MaxEnt` reward function to the L_1-Coverage reward function (Line 4 in Algorithm 1).

Environment. We evaluate on the `MountainCarContinuous-v0` environment (henceforth simply `MountainCar`) from the OpenAI Gym (Brockman et al., 2016). The state space of the environment is two-dimensional, with a position value (denoted by ξ) that is in the interval $[-1.2, 0.6]$ and a velocity value (denoted by ρ) that is in the interval $[-0.07, 0.07]$. The dynamics are deterministic and defined by the physics of a sinusoidal valley, we refer the reader to the documentation for the precise equations.[12] The goal state (the "flag") is at the top of the right hill, with a position $\xi = +0.45$. The bottom of the valley corresponds to the coordinate $\xi = -\pi/6 \approx -0.52$. We modify the environment to have a deterministic starting state with a position of $\xi = -0.5$ and a velocity of $\rho = 0$, so that more deliberate exploration is required to find a high-quality cover. This means that occupancies and covers are evaluated when rolling out from this deterministic starting state. The action space is continuous in the interval $[-1, 1]$, with negative values corresponding to forces applied to the left and positive values corresponding to forces applied to the right. To simplify, we will consider that the environment only has 3 actions, namely we only allow actions in $\{-1, 0, 1\}$. Finally, we take a horizon of $H = 200$, meaning that we terminate rollouts after 200 steps (if the goal state has not been reached yet).

Implementation of Algorithm 1. We make a few changes to Algorithm 1. Firstly, while `MountainCar` has a time horizon and is thus non-stationary, we approximate the dynamics as stationary. Thus, we replace $d_h^{M,\pi}$ in Line 4 of Algorithm 1 by a stationary analogue $d_{\mathtt{stat}}^{M,\pi}$, and only invoke Algorithm 1 once rather than at each layer h. Namely, we define

$$d_{\mathtt{stat}}^{M,\pi} := \frac{1}{H} \sum_h d_h^{M,\pi},$$

where $H = 200$ is the horizon which we define for `MountainCar`. We will henceforth simply write $d^\pi := d_{\mathtt{stat}}^{M,\pi}$ for compactness. Similarly, we choose a stationary distribution μ (to be described shortly) to pass as input to the algorithm. Secondly, rather than solving the policy optimization problem for the reward function

$$r(x, a) = \frac{\mu(x, a)}{\sum_{i<t} d^{M,\pi^i}(x, a) + C_\infty \mu(x, a)}$$

as written Line 4, we instead apply some regularization and solve the policy optimization problem for the more general L_1-Coverage-based reward function defined in Eq. (6), namely

$$r(x, a) = \frac{\mu(x, a)}{\sum_{i<t} d^{M,\pi^i}(x, a) + \varepsilon C_\infty \mu(x, a)}, \tag{16}$$

for a small parameter $\varepsilon > 0$. This has the effect of magnifying the reward difference between visited and unvisited states. We also (linearly) renormalize so that this reward function lies in the interval $[0, 1]$. We found that regularizing with ε helps our approximate planning subroutine (discussed shortly) solve the policy optimization and recover a better policy for the L_1-Coverage-based reward function.

[12] https://www.gymlibrary.dev/environments/classic_control/mountain_car_continuous/

Approximating d^π and μ. To approximate d^π and choose μ, we define a discretized state space for `MountainCar`. Namely, we discretize the position space $[-1.2, 0.6]$ in 12 evenly-spaced intervals and the velocity space $[-0.07, 0.07]$ in 11 evenly-spaced intervals. This gives a tabular discretization with $12 \times 11 = 132$ states. We then approximate $d^\pi(x, a)$ via a simple count-based estimate on the discretized space. That is, letting the discretized space be denoted by $\mathcal{B} = \{b_1, \ldots, b_{132}\}$ where each b_i is a bin, and given trajectories $(x_h^n, a_h^n)_{n\in[N],h\in[H]}$, we estimate

$$d^\pi_{\texttt{disc}}(b_i, a) \coloneqq \frac{1}{H}\sum_{h=1}^{H} d^\pi_{\texttt{disc},h}(b_i, a), \quad \text{where} \quad d^\pi_{\texttt{disc},h}(b_i, a) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\{x_h^n \in b_i, a_h^n = a\}$$

and then for any x, a we assign

$$d^\pi(x, a) = d^\pi_{\texttt{disc}}(b_{i_x}, a),$$

where i_x is the index of the bin b_i in which state x lies. For the distribution μ and the L_∞ coverability parameter C_∞, we take μ to be uniform over the discretized state-action space, that is we take

$$\mu(x, a) = \frac{1}{|\mathcal{B}||\mathcal{A}|} = \frac{1}{396},$$

and

$$C_\infty = |\mathcal{B}||\mathcal{A}| = 396.$$

Approximate planner. As an approximate planner, we take a simple implementation of the `REINFORCE` algorithm (Sutton et al., 1999) given in the PyTorch (Paszke et al., 2019) GitHub repository[13], which only differs from the classical `REINFORCE` in that it applies variance-smoothing (with some parameter σ) to the returns. When solving the policy optimization problem, we allow `REINFORCE` to use the original stochastic reset distribution for `MountainCar`, that is the reset distribution which samples $\xi \in [-0.6, -0.4]$ uniformly and has $\rho = 0$.

Architecture, optimizer, hyperparameters. The policy class we use, and which `REINFORCE` optimizes over, is obtained by a set of fully-connected feedforward neural nets with ReLU activation, 1 hidden layer, and 128 hidden units. The input is 2-dimensional (corresponding to the 2-dimensional state space of `MountainCar`) and the output is a 3-dimensional vector; we obtain a distribution over the action set $\{-1, 0, 1\}$ by taking a softmax over the output. We use Xavier initialization (Glorot & Bengio, 2010).

For `REINFORCE`, we take a discount factor of 0.99, and a variance smoothing parameter of $\sigma = 0.05$. We train `REINFORCE` with horizons of length 400. We take π^t, the policy which approximates Line 4 of Algorithm 1, to be the policy returned after 1000 `REINFORCE` updates, with one update after each rollout. The update in `REINFORCE` use the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 10^{-3}.

We estimate all occupancies with $N = 100$ rollouts of length $H = 200$. We calculate the mixture occupancies d^p, for $p = \mathrm{Unif}(\pi^1, \ldots, \pi^t)$, by estimating the occupancy for each d^{π_i} separately and averaging via

$$d^p(x, a) = \frac{1}{t}\sum_{i=1}^{t} d^{\pi^i}(x, a).$$

We train for 20 epochs, corresponding to $T = 20$ in the loop of Line 3 of Algorithm 1. For the regularized reward of Eq. (16), we take $\varepsilon = 10^{-4}$.

B.2. Additional Experimental Results

B.2.1. MOUNTAINCAR

In addition to the results of Section 6, in Figure 2 we report the entropy and L_1-Coverability of the state distributions for the policy covers found by the three algorithms throughout training. Namely, for each cover p, we estimate its occupancy d^p via the procedure defined above, and measure the entropy of our estimate for d^p. In the second plot, we take the estimate of d^p

[13] https://github.com/pytorch/examples/blob/main/reinforcement_learning/reinforce.py

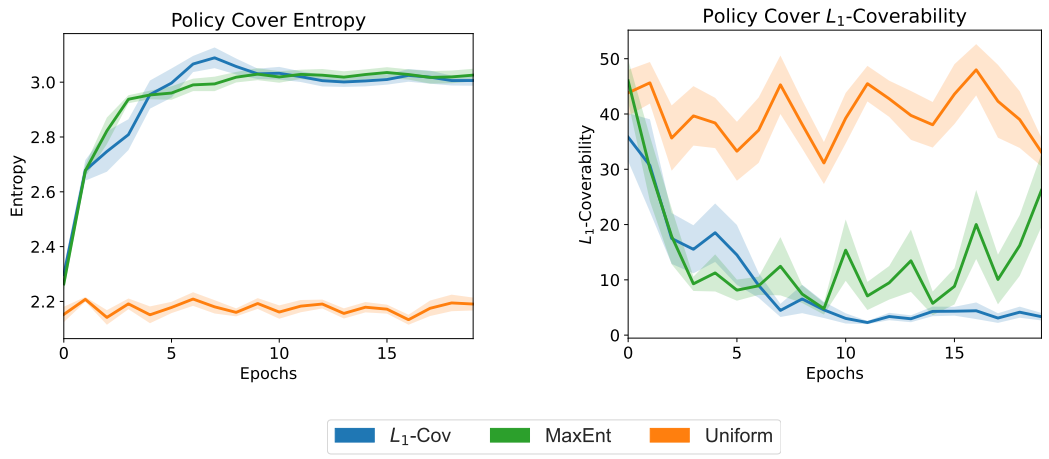

Figure 2. Entropy and L_1-Coverability measured on each policy cover obtained from L_1-Coverage (Algorithm 1), `MaxEnt`, and uniform exploration on the `MountainCar` environment. We plot the mean and standard error across 10 runs. Each epoch corresponds to a single policy update in Algorithm 1 and `MaxEnt`, obtained through 1000 steps of `REINFORCE` with rollouts of length 400.

and measure its objective value $\Psi^M_{\mu;h,\varepsilon}(p)$, where $\Psi^M_{\mu;h,\varepsilon}(p)$ is the L_∞-Coverability relaxation of L_1-Coverability which we recall is defined via

$$\Psi^M_{\mu;h,\varepsilon}(p) = \sup_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d^{M,p}_h(x_h,a_h) + \varepsilon \cdot C_\infty \mu(x_h,a_h)}\right]. \tag{17}$$

We measure this quantity with $\varepsilon = 10^{-4}$, and in the same way that we approximate Line 4 in Algorithm 1, namely by calling `REINFORCE` with a stochastic starting distribution as an approximate planner, with the same parameters and architecture. We note that there are two sources of approximation error in our measurements of $\Psi^M_{\mu;h,\varepsilon}(p)$, namely the estimation error of d^p (via a count-based estimate on the discretized space) and the optimization error of `REINFORCE`, and thus values we report should be taken as approximations of the true L_1-Coverability values.[14]

Results. We find that L_1-Coverage and `MaxEnt` recover policy covers with similar entropy values, with the entropy of the `MaxEnt` cover being slightly larger (despite the `MaxEnt` cover visiting fewer states, as seen in our results in Section 6). Interestingly, the `MaxEnt` baseline has higher entropy while visiting fewer states (as seen in our results in Section 6), indicating that entropy may not be the best proxy for exploration. For L_1-Coverage, we find that Algorithm 1 attains the smallest L_1-Coverage values, indicating a better cover.

B.2.2. Pendulum

Environment. We evaluate on the `Pendulum-v0` environment (henceforth simply `Pendulum`) from the OpenAI Gym (Brockman et al., 2016). The state space of the environment is two-dimensional, with an angle value (denoted by θ) that is in the interval $[-\pi,\pi]$ and a velocity value (denoted by ρ) that is in the interval $[-8,8]$. The dynamics are deterministic and defined by the physics of an inverted pendulum. We modify the starting distribution to be a deterministic starting state with a position of $\theta = \pi$ and a velocity of $\rho = 0$. This means that occupancies and covers are evaluated when rolling out from this deterministic starting state. The action space is continuous in the interval $[-2,2]$, with negative values corresponding to torque applied to the left and positive values corresponding to torque applied to the right. To simplify, we only allow actions in $\{-1,0,1\}$, so that $|\mathcal{A}| = 3$. Finally, we take a horizon of $H = 200$, meaning that we terminate rollouts after 200 steps.

Discretization and other hyperparameters. We apply all the same implementation details as in Appendix B.1. We take a discretization resolution for `Pendulum` of 8×8. We take a `REINFORCE` planner with a uniform starting distribution with $\theta \in [-\pi,\pi]$ and $\rho \in [-1,1]$. The architecture, optimizer, and hyperparameters are the same as for `MountainCar`.

[14] For instance, notice the variability over epochs of the uniform random baseline, which has a constant $\Psi^M_{\mu;h,\varepsilon}(p)$ value.

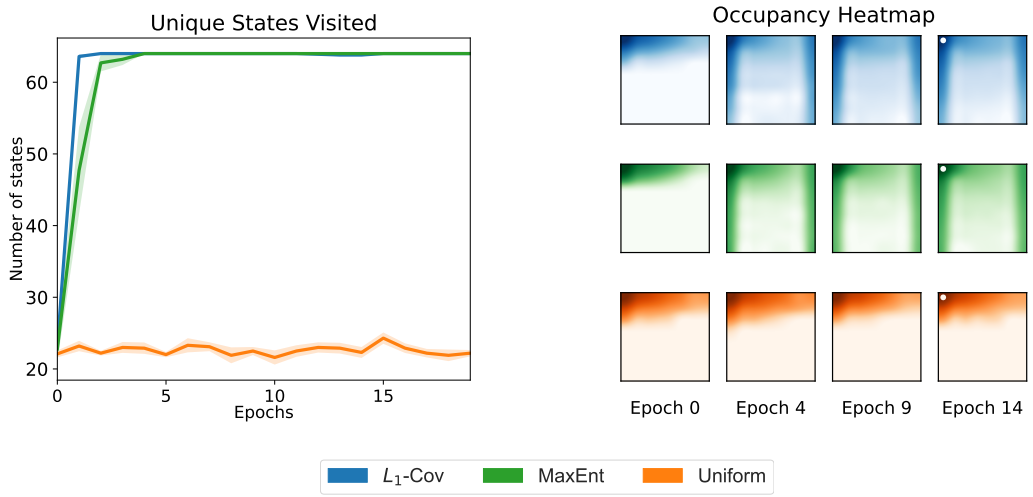

Figure 3. Number of discrete states visited (mean and standard error over 10 runs) and occupancy heatmaps for each policy cover obtained from L_1-Coverage (Algorithm 1), `MaxEnt`, and uniform exploration in the `Pendulum` environment. Each epoch comprises a single policy update in Algorithm 1 and `MaxEnt`, obtained through 1000 steps of `REINFORCE` with rollouts of length 400. Heatmap axes: torque (x-axis) and angle (y-axis). Start state indicated by •.

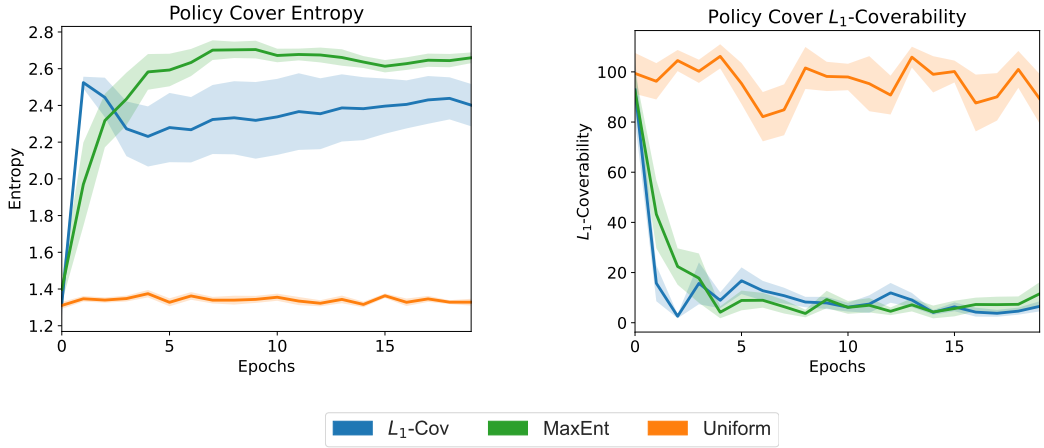

Figure 4. Entropy and L_1-Coverability measured on each policy cover obtained from L_1-Coverage (Algorithm 1), `MaxEnt`, and uniform exploration on the `Pendulum` environment. We plot the mean and standard error across 10 runs. Each epoch corresponds to a single policy update in Algorithm 1 and `MaxEnt`, obtained through 1000 steps of `REINFORCE` with rollouts of length 400.

Results. As with MountainCar, we measure the number of unique states visited and the occupancy heatmaps (Figure 3) as well as the entropy and L_1-Coverage values (Figure 4). Overall, we find that L_1-Coverage and MaxEnt obtain similar values and are able to explore every state in the (discretized) Pendulum environment, indicating that exploration is somewhat easier than for the MountainCar environment.

B.3. Additional Discussion

While the MountainCar and Pendulum environments are fairly simple, we note that our algorithm exhibits robustness in several different ways, indicating that it may scale favorably to more challenging domains. Firstly, we do not expect that REINFORCE is finding a near-optimal policy for the planning problem in Line 4, which indicates that our method is robust to optimization errors. We also note that the choice of μ used in our implementation of Algorithm 1, which is defined as uniform over the discretized space, is a heuristic and is not guaranteed to have good coverage with respect to the true continuous MDP. This indicates that our method is robust to the choice of μ and C_∞. Lastly, our method worked "out-of-the-box" with the same hyperparameters used in the MaxEnt implementation in Hazan et al. (2019), and was found to behave similarly with different hyperparameters, which indicates that our method is robust to the choice of hyperparameters.

C. Technical Tools

C.1. Minimax Theorem

Lemma C.1 (Sion's Minimax Theorem (Sion, 1958))**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be convex sets in linear topological spaces, and assume $\mathcal{X}$ is compact. Let $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be such that (i) $f(x, \cdot)$ is concave and upper semicontinuous over $\mathcal{Y}$ for all $x \in \mathcal{X}$ and (ii) $f(\cdot, y)$ is convex and lower semicontinuous over $\mathcal{X}$ for all $y \in \mathcal{Y}$. Then*

$$\inf_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} f(x, y) = \sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} f(x, y). \tag{18}$$

C.2. Concentration

Lemma C.2. *For any sequence of real-valued random variables $(X_t)_{t\le T}$ adapted to a filtration $(\mathscr{F}_t)_{t\le T}$, it holds that with probability at least $1 - \delta$, for all $T' \le T$,*

$$\sum_{t=1}^{T'} -\log\big(\mathbb{E}_{t-1}\big[e^{-X_t}\big]\big) \le \sum_{t=1}^{T'} X_t + \log(\delta^{-1}). \tag{19}$$

Lemma C.3 (Freedman's inequality (e.g., Agarwal et al. (2014)))**.** *Let $(X_t)_{t\le T}$ be a real-valued martingale difference sequence adapted to a filtration $(\mathscr{F}_t)_{t\le T}$. If $|X_t| \le R$ almost surely, then for any $\eta \in (0, 1/R)$, with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} X_t \le \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}\big[X_t^2\big] + \frac{\log(\delta^{-1})}{\eta}.$$

Lemma C.4 (Corollary of Lemma C.3)**.** *Let $(X_t)_{t\le T}$ be a sequence of random variables adapted to a filtration $(\mathscr{F}_t)_{t\le T}$. If $0 \le X_t \le R$ almost surely, then with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} X_t \le \frac{3}{2} \sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] + 4R \log(2\delta^{-1}),$$

and

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] \le 2 \sum_{t=1}^{T} X_t + 8R \log(2\delta^{-1}).$$

C.3. Information Theory

Lemma C.5 (e.g., Foster et al. (2021))**.** *For any pair of random variables (X, Y),*

$$\mathbb{E}_{X\sim\mathbb{P}_X}\big[D^2_{\mathsf{H}}\big(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\big)\big] \le 4D^2_{\mathsf{H}}(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}).$$

Lemma C.6 (Foster et al. (2021))**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $(\mathcal{X}, \mathscr{F})$. For all $h : \mathcal{X} \to \mathbb{R}$ with $0 \le h(X) \le B$ almost surely under $\mathbb{P}$ and $\mathbb{Q}$, we have*

$$\mathbb{E}_{\mathbb{P}}[h(X)] \le 3\,\mathbb{E}_{\mathbb{Q}}[h(X)] + 4B \cdot D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}). \tag{20}$$

C.4. Reinforcement Learning

Proofs for the following lemmas can be found in Foster et al. (2021).

Lemma C.7 (Global simulation lemma)**.** *Let M and M' be MDPs with $\sum_{h=1}^{H} r_h \in [0, 1]$ almost surely, and let $\pi \in \Pi_{\mathrm{RNS}}$. Then we have*

$$\big|f^M(\pi) - f^{M'}(\pi)\big| \le D_{\mathsf{H}}(M(\pi), M'(\pi)) \le \frac{1}{2\eta} + \frac{\eta}{2} D^2_{\mathsf{H}}(M(\pi), M'(\pi)) \quad \forall \eta > 0. \tag{21}$$

Lemma C.8 (Local simulation lemma)**.** *For any pair of MDPs $M = (P^M, R^M)$ and $\overline{M} = (P^{\overline{M}}, R^{\overline{M}})$ with the same initial state distribution and $\sum_{h=1}^{H} r_h \in [0, 1]$,*

$$\big|f^M(\pi) - f^{\overline{M}}(\pi)\big| \le \sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi}\big[D_{\mathsf{H}}\big(P^M_h(x_h, a_h), P^{\overline{M}}_h(x_h, a_h)\big) + D_{\mathsf{H}}\big(R^M_h(x_h, a_h), R^{\overline{M}}_h(x_h, a_h)\big)\big]. \tag{22}$$

C.5. Helper Lemmas

Lemma C.9 (E.g., Xie et al. (2023))**.** *Let $d^1, d^2, \ldots, d^T$ be an arbitrary sequence of distributions over a set $\mathcal{Z}$, and let $\mu \in \Delta(\mathcal{Z})$ be a distribution such that $d^t(z)/\mu(z) \le C$ for all $(z,t) \in \mathcal{Z} \times [T]$. Then for all $z \in \mathcal{Z}$, we have*

$$\sum_{t=1}^{T} \frac{d^t(z)}{\sum_{i<t} d^i(z) + C \cdot \mu(z)} \le 2\log(2T).$$

Lemma C.10. *For any distribution $p \in \Delta(\Pi_{\mathsf{rns}})$, $\pi \in \Pi_{\mathsf{rns}}$, $\mu \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{A}}$, and $\varepsilon, \delta > 0$, we have that*

$$\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon \cdot d_h^{M,\pi}(x_h,a_h)}\right] \le \mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \delta \cdot \mu(x_h,a_h)}\right] + \frac{\delta}{\varepsilon} \cdot \mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \delta \cdot \mu(x_h,a_h)}\right].$$

Proof of Lemma C.10. The result follows by observing that we can bound

$$\begin{aligned}
&\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon \cdot d_h^{M,\pi}(x_h,a_h)}\right] - \mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \delta \cdot \mu(x_h,a_h)}\right] \\
&= \sum_{x\in\mathcal{X},a\in\mathcal{A}} \frac{(d_h^{M,\pi}(x,a))^2\big(\delta \cdot \mu(x,a) - \varepsilon \cdot d_h^{M,\pi}(x,a)\big)}{(d_h^{M,p}(x,a) + \varepsilon \cdot d_h^{M,\pi}(x,a))(d_h^{M,p}(x,a) + \delta \cdot \mu(x,a))} \\
&\le \sum_{x\in\mathcal{X},a\in\mathcal{A}} \frac{(d_h^{M,\pi}(x,a))^2(\delta \cdot \mu(x,a))}{(d_h^{M,p}(x,a) + \varepsilon \cdot d_h^{M,\pi}(x,a))(d_h^{M,p}(x,a) + \delta \cdot \mu(x,a))} \\
&\le \frac{\delta}{\varepsilon} \sum_{x\in\mathcal{X},a\in\mathcal{A}} \frac{d_h^{M,\pi}(x,a)\mu(x,a)}{d_h^{M,p}(x,a) + \delta \cdot \mu(x,a)}.
\end{aligned}$$

□

Lemma C.11. *For any $\pi \in \Pi_{\mathsf{rns}}$, $d \in \mathbb{R}_+^{\mathcal{X}}$, $\mu \in \mathbb{R}_+^{\mathcal{X}}$, and $\varepsilon, \delta > 0$, we have that*

$$\mathbb{E}^{M,\pi}\left[\frac{P_{h-1}^M(x_h \mid x_{h-1},a_{h-1})}{d(x_h) + \varepsilon \cdot P_{h-1}^M(x_h \mid x_{h-1},a_{h-1})}\right] \le \mathbb{E}^{M,\pi}\left[\frac{P_{h-1}^M(x_h \mid x_{h-1},a_{h-1})}{d(x_h) + \delta \cdot \mu(x_h)}\right] + \frac{\delta}{\varepsilon} \cdot \mathbb{E}^{M,\pi}\left[\frac{\mu(x_h)}{d(x_h) + \delta \cdot \mu(x_h)}\right].$$

Proof of Lemma C.11. The result follows using similar reasoning to Lemma C.10:

$$\begin{aligned}
&\mathbb{E}^{M,\pi}\left[\frac{P_{h-1}^M(x_h \mid x_{h-1},a_{h-1})}{d(x_h) + \varepsilon \cdot P_{h-1}^M(x_h \mid x_{h-1},a_{h-1})}\right] - \mathbb{E}^{M,\pi}\left[\frac{P_{h-1}^M(x_h \mid x_{h-1},a_{h-1})}{d(x_h) + \delta \cdot \mu(x_h)}\right] \\
&= \sum_{x\in\mathcal{X},a\in\mathcal{A},x'\in\mathcal{X}} d_{h-1}^{M,\pi}(x,a) \frac{(P_{h-1}^M(x' \mid x,a))^2\big(\delta \cdot \mu(x') - \varepsilon \cdot P_{h-1}^M(x' \mid x,a)\big)}{(d(x') + \varepsilon \cdot P_{h-1}^{M,}(x' \mid x,a))(d(x') + \delta \cdot \mu(x'))} \\
&\le \sum_{x\in\mathcal{X},a\in\mathcal{A},x'\in\mathcal{X}} d_{h-1}^{M,\pi}(x,a) \frac{(P_{h-1}^M(x' \mid x,a))^2(\delta \cdot \mu(x'))}{(d(x') + \varepsilon \cdot P_{h-1}^{M,}(x' \mid x,a))(d(x') + \delta \cdot \mu(x'))} \\
&\le \frac{\delta}{\varepsilon} \sum_{x\in\mathcal{X},a\in\mathcal{A},x'\in\mathcal{X}} d_{h-1}^{M,\pi}(x,a) \frac{P_{h-1}^M(x' \mid x,a)\mu(x')}{d(x') + \delta \cdot \mu(x')} \\
&= \frac{\delta}{\varepsilon} \cdot \mathbb{E}^{M,\pi}\left[\frac{\mu(x_h)}{d(x_h) + \delta \cdot \mu(x_h)}\right].
\end{aligned}$$

□

Part I

Additional Results and Discussion

D. L_1-Coverage: Application to Downstream Policy Optimization

In this section, we show how to use data gathered using policy covers with bounded L_1-Coverage to perform offline policy optimization. Here, there is an underlying (reward-free) MDP $M^\star = \{\mathcal{X}, \mathcal{A}, \{P_h^{M^\star}\}_{h=0}^H\}$ and a given policy class Π, and we have access to policy covers $p_1, \ldots, p_H$ such that the L_1-Coverage objective $\Psi_{h,\varepsilon}^{M^\star}(p_h)$ is small for all h. For an arbitrary user-specified reward distribution $R = \{R_h\}_{h=1}^H$ with $\sum_{h=1}^H r_h \in [0,1]$ almost surely, we define

$$J_R^{M^\star}(\pi) := \mathbb{E}^{M^\star,\pi}\left[\sum_{h=1}^H r_h\right]$$

as the value under $r_h \sim R(x_h, a_h)$. Our goal is to use trajectories drawn from $p_1, \ldots, p_H$ to compute a policy $\widehat{\pi}$ such

$$\max_{\pi\in\Pi} J_R^{M^\star}(\pi) - J_R^{M^\star}(\widehat{\pi}) \le \epsilon$$

using as few samples as possible.

We present guarantees for two standard offline policy optimization methods: Maximum Likelihood Estimation (MLE) and Fitted Q-Iteration (FQI). While both of these algorithms are fully offline, combining them with the online algorithms for reward-free exploration in Section 5 leads to end-to-end algorithms for online reward-driven exploration. We expect that similar guarantees can be proven for other standard offline policy optimization methods (Munos, 2003; Antos et al., 2008; Chen & Jiang, 2019; Xie & Jiang, 2020; 2021; Jin et al., 2021b; Rashidinejad et al., 2021; Foster et al., 2022; Zhan et al., 2022), as well as offline policy evaluation methods (Liu et al., 2018; Uehara et al., 2020; Yang et al., 2020; Uehara et al., 2021) and hybrid offline/online methods (Bagnell et al., 2003; Agarwal et al., 2021; Song et al., 2022).

Maximum likelihood. Let $\mathcal{M}$ be a realizable model class for which $M^\star \in \mathcal{M}$. We define the Maximum Likelihood algorithm as follows:

- For each $h \in [H]$, gather n trajectories $\{o^{h,t}\}_{t\in[n]}$, where $o^{h,t} = (x_1^{h,t}, a_1^{h,t}, r_1^{h,t}), \ldots, (x_H^{h,t}, a_H^{h,t}, r_H^{h,t})$ by executing $\pi^{h,t} \sim p_h$ in $M^\star$ with R as the reward distribution.
- Set $\widehat{M} = \arg\max_{M\in\mathcal{M}} \sum_{h=1}^H \sum_{t=1}^n \log(M(o^{h,t} \mid \pi^{h,t}))$, where $M(o \mid \pi)$ denotes the likelihood of trajectory o under π in M.
- Let $\widehat{\pi} = \arg\max_{\pi\in\Pi} J_R^{M^\star}(\pi)$.

Proposition D.1. *Assume that $M^\star \in \mathcal{M}$ and let $R = \{R_h\}_{h=1}^H$ be a reward distribution with $\sum_{h=1}^H r_h \in [0,1]$ almost surely for all $M \in \mathcal{M}$. For any $n \in \mathbb{N}$, given policy covers $p_1, \ldots, p_H$, the Maximum Likelihood algorithm ensures that with probability at least $1-\delta$,*

$$J_R^{M^\star}(\pi_\star) - J_R^{M^\star}(\widehat{\pi}) \le 8H\left(\sqrt{\max_h \Psi_{h,\varepsilon}^{M^\star}(p_h) \cdot \frac{\log(|\mathcal{M}|/\delta)}{n}} + \max_h \Psi_{h,\varepsilon}^{M^\star}(p_h)\cdot\varepsilon\right), \tag{23}$$

and uses $H \cdot n$ trajectories total.

Fitted Q-iteration. Let a value function class $\mathcal{Q} = \mathcal{Q}_1, \ldots, \mathcal{Q}_H \subset (\mathcal{X}\times\mathcal{A}) \to [0,1]$ be given. For a value function Q, define the Bellman operator for $M^\star$ with reward distribution R by $[\mathcal{T}_{R,h}^{M^\star} Q](x,a) := \mathbb{E}^{M^\star}[r_h + \max_{a'\in\mathcal{A}} Q(x_{h+1}, a') \mid x_h = x, a_h = a]$ under $r_h \sim R(x_h, a_h)$. We make the Bellman completeness assumption that for all $Q \in \mathcal{Q}_{h+1}$, $[\mathcal{T}_{R,h}^{M^\star} Q] \in \mathcal{Q}_h$. We define the L_1-Coverage objective with respect to the policy class $\Pi = \{\pi_Q\}_{Q\in\mathcal{Q}}$, where $\pi_{Q,h}(x) := \arg\max_{a\in\mathcal{A}} Q_h(x,a)$.

The Fitted Q-Iteration algorithm is defined as follows:

- For each $h = H, \dots, 1$:
  - Gather n trajectories $\{o^{h,t}\}_{t\in[n]}$, where $o^{h,t} = (x_1^{h,t}, a_1^{h,t}, r_1^{h,t}), \dots, (x_H^{h,t}, a_H^{h,t}, r_H^{h,t})$ by executing $\pi^{h,t} \sim p_h$ in $M^\star$ with R as the reward distribution.
  - Set $\widehat{Q}_h = \arg\min_{Q\in\mathcal{Q}_h} \sum_{t=1}^n \big(Q(x_h^{h,t}, a_h^{h,t}) - r_h^{h,t} - \max_{a'\in\mathcal{A}} \widehat{Q}_{h+1}(x_{h+1}^{h,t}, a')\big)^2$.
- Let $\widehat{\pi}_h(x) := \arg\max_{a\in\mathcal{A}} \widehat{Q}_h(x, a)$.

Proposition D.2. *Let $R = \{R_h\}_{h=1}^H$ be a reward distribution with $r_h \in [0,1]$ and $\sum_{h=1}^H r_h \in [0,1]$ almost surely, and assume that for all $Q \in \mathcal{Q}_{h+1}$, $\big[\mathcal{T}_{R,h}^{M^\star} Q\big] \in \mathcal{Q}_h$. For any $n \in \mathbb{N}$, given policy covers $p_1, \dots, p_H$, the Fitted Q-Iteration algorithm ensures that with probability at least $1-\delta$,*

$$J_R^{M^\star}(\pi_\star) - J_R^{M^\star}(\widehat{\pi}) \le O(H) \cdot \left(\sqrt{\max_h \Psi_{h,\varepsilon}^{M^\star}(p_h) \cdot \frac{\log(|\mathcal{Q}|H/\delta)}{n}} + \max_h \Psi_{h,\varepsilon}^{M^\star}(p_h) \cdot \varepsilon \right), \tag{24}$$

and uses $H \cdot n$ trajectories total.

D.1. Proofs

Proof of Proposition D.1. By the standard generalization bound for MLE (e.g., Foster & Krishnamurthy (2021); Foster & Rakhlin (2023)), we are guaranteed that with probability at least $1-\delta$,

$$\sum_{h=1}^H \mathbb{E}^{M^\star, p_h}\big[D_{\mathsf{H}}^2\big(P_h^{\widehat{M}}(x_h, a_h), P_h^{M^\star}(x_h, a_h)\big)\big] \le \sum_{h=1}^H \mathbb{E}_{\pi\sim p_h}\Big[D_{\mathsf{H}}^2\Big(\widehat{M}(\pi), M^\star(\pi)\Big)\Big] \le 2\frac{\log(|\mathcal{M}|/\delta)}{n}.$$

Let $\pi_\star := \arg\max_{\pi\in\Pi} J_R^{M^\star}(\pi)$. Using Lemma C.8, we have that

$$J_R^{M^\star}(\pi_\star) - J_R^{M^\star}(\widehat{\pi}) \le J_R^{M^\star}(\pi_\star) - J_R^{\widehat{M}}(\pi_\star) + J_R^{\widehat{M}}(\widehat{\pi}) - J_R^{M^\star}(\widehat{\pi}) \le 2\max_{\pi\in\Pi} \sum_{h=1}^H \mathbb{E}^{M^\star,\pi}\big[D_{\mathsf{H}}\big(P_h^{\widehat{M}}(x_h, a_h), P_h^{M^\star}(x_h, a_h)\big)\big].$$

Since $D_{\mathsf{H}}\big(P_h^{\widehat{M}}(x_h, a_h), P_h^{M^\star}(x_h, a_h)\big) \in [0, \sqrt{2}]$, we can use Proposition 3.1 to bound

$$\begin{aligned}
\max_{\pi\in\Pi} \mathbb{E}^{M^\star,\pi}\big[D_{\mathsf{H}}\big(P_h^{\widehat{M}}(x_h, a_h), P_h^{M^\star}(x_h, a_h)\big)\big] &\le 2\sqrt{\Psi_{h,\varepsilon}^{M^\star}(p_h) \cdot \mathbb{E}^{M^\star,p_h}\big[D_{\mathsf{H}}^2\big(P_h^{\widehat{M}}(x_h, a_h), P_h^{M^\star}(x_h, a_h)\big)\big]} \\
&\quad + \sqrt{2}\Psi_{h,\varepsilon}^{M^\star}(p_h)\varepsilon \\
&\le 2\sqrt{2\Psi_{h,\varepsilon}^{M^\star}(p_h) \cdot \frac{\log(|\mathcal{M}|/\delta)}{n}} + \sqrt{2} \cdot \Psi_{h,\varepsilon}^{M^\star}(p_h) \cdot \varepsilon.
\end{aligned}$$

Combining this with the preceding inequalities yields the result.

□

Proof of Proposition D.2. By a standard generalization bound for FQI (e.g., Xie & Jiang (2020); Xie et al. (2023)), under the Bellman completeness assumption, it holds that with probability at least $1-\delta$,

$$\sum_{h=1}^H \mathbb{E}^{M^\star,p_h}\left[\Big(\widehat{Q}_h(x_h, a_h) - \big[\mathcal{T}_{R,h}^{M^\star}\widehat{Q}_{h+1}\big](x_h, a_h)\Big)^2\right] \le O\Big(\frac{\log(|\mathcal{Q}|H/\delta)}{n}\Big).$$

At the same time, using a finite-horizon adaptation of Xie & Jiang (2020, Lemma 4), we have that

$$J_R^{M^\star}(\pi_\star) - J_R^{M^\star}(\widehat{\pi}) \le 2\max_{\pi\in\Pi} \sum_{h=1}^H \mathbb{E}^{M^\star,\pi}\Big[\Big|\widehat{Q}_h(x_h, a_h) - \big[\mathcal{T}_{R,h}^{M^\star}\widehat{Q}_{h+1}\big](x_h, a_h)\Big|\Big].$$

Since $\left|\widehat{Q}_h(x_h, a_h) - \left[\mathcal{T}^{M^\star}_{R,h}\widehat{Q}_{h+1}\right](x_h, a_h)\right| \in [0, 1]$, we can use Proposition 3.1 to bound

$$\begin{aligned}
&\max_{\pi\in\Pi} \mathbb{E}^{M^\star,\pi}\left[\left|\widehat{Q}_h(x_h, a_h) - \left[\mathcal{T}^{M^\star}_{R,h}\widehat{Q}_{h+1}\right](x_h, a_h)\right|\right] \\
&\leq 2\sqrt{\Psi^{M^\star}_{h,\varepsilon}(p_h) \cdot \mathbb{E}^{M^\star,p_h}\left[\left(\widehat{Q}_h(x_h, a_h) - \left[\mathcal{T}^{M^\star}_{R,h}\widehat{Q}_{h+1}\right](x_h, a_h)\right)^2\right]} + \Psi^{M^\star}_{h,\varepsilon}(p_h) \cdot \varepsilon \\
&\leq O\left(\sqrt{\Psi^{M^\star}_{h,\varepsilon}(p_h) \cdot \frac{\log(|\mathcal{Q}|H/\delta)}{n}} + \Psi^{M^\star}_{h,\varepsilon}(p_h) \cdot \varepsilon\right).
\end{aligned}$$

□

E. Efficient Model-Free Exploration via L_1-Coverage

Our algorithms in the previous section show that the L_1-Coverage objective and L_1-Coverability parameter enable sample-efficient online reinforcement learning, but one potential drawback is that they require model-based realizability, a strong form of function approximation that may not always be realistic. In this section, we give *model-free* algorithms to perform reward-free exploration and optimize the L_1-Coverage objective that do not require model realizability, and instead require a weaker form of *density ratio* or *weight function* realizability, a modeling approach that has been widely used in offline reinforcement learning (Liu et al., 2018; Uehara et al., 2020; Yang et al., 2020; Uehara et al., 2021; Jiang & Huang, 2020; Xie & Jiang, 2020; Zhan et al., 2022; Chen & Jiang, 2022; Rashidinejad et al., 2023; Ozdaglar et al., 2023) and recently adapted to the online setting (Amortila et al., 2024). The main algorithm we present computes a policy cover that achieves a bound on the L_1-Coverability objective that scales with the pushforward coverability parameter (Section 4.2), but the weaker modeling assumptions make it applicable in a broader range of settings.

Throughout this section, we take $\Pi = \Pi_{\mathsf{ns}}$ as the set of all non-stationary policies.

E.1. Algorithm

Our model-free algorithm, `CODEX.W`, is presented in Algorithm 3. The algorithm builds a collection of policy covers $p_1, \ldots, p_H \in \Delta(\Pi_{\mathsf{rns}})$ layer-by-layer in an inductive fashion. For each layer $h \in [H]$, given policy covers $p_1, \ldots, p_{h-1}$ for the preceding layers, the algorithm computes p_h by (approximately) implementing the iterative algorithm for policy cover construction given in Algorithm 5 (Section 4.2), in a data-driven fashion.[15]

In more detail, recall the *pushforward coverability* relaxation

$$\Psi^{M^\star}_{\mathsf{push};h,\varepsilon}(p) = \sup_{\pi \in \Pi} \mathbb{E}^{M^\star,\pi}\left[\frac{P^{M^\star}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{d^{M^\star,p}_h(x_h) + \varepsilon \cdot P^{M^\star}_{h-1}(x_h \mid x_{h-1}a_{h-1})}\right]$$

for the L_1-Coverage objective given in Section 4.2. For layer h, Algorithm 3 approximately minimizes this objective by computing a sequence of policies $\pi^{h,1}, \ldots, \pi^{h,T}$, where each policy

$$\pi^{h,t} \approx \arg\max_{\pi \in \Pi} \mathbb{E}^{M^\star,\pi}\left[\frac{P^{M^\star}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\sum_{i<t} d^{\pi^{h,i}}_h(x_h) + P^{M^\star}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\right] \tag{25}$$

is computed in a data-driven, online fashion that makes use of the preceding policy covers $p_1, \ldots, p_{h-1}$. The algorithm then computes the cover p_h via $p_h = \mathrm{Unif}(\pi^{h,1} \circ_h \pi_{\mathrm{unif}}, \ldots, \pi^{h,T} \circ_h \pi_{\mathrm{unif}})$.

Our planning analysis in Section 4.2 shows that as long as the approximation error in Eq. (25) is small, p_h will indeed be an approximate policy cover that minimizes $\Psi^{M^\star}_{\mathsf{push};h,\varepsilon}(p_h)$. To achieve this, Algorithm 3 makes use of two subroutines. The first subroutine, `EstimateWeightFunction` (Algorithm 4), invoked in Line 6, uses function approximation to estimate a weight function $\widehat{w}^t_h$ such that

$$\widehat{w}^t_h(x_h \mid x_{h-1}, a_{h-1}) \approx w^t_h(x_h \mid x_{h-1}, a_{h-1}) := \frac{P^{M^\star}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\sum_{i<t} d^{\pi^{h,i}}_h(x_h) + P^{M^\star}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}. \tag{26}$$

The second subroutine, `PolicyOptimization`, is a hyperparameter to the algorithm, and approximately solves the policy optimization problem

$$\pi^{h,t} \approx \arg\max_{\pi \in \Pi} \mathbb{E}^{M^\star,\pi}[\widehat{w}^t_h(x_h \mid x_{h-1}, a_{h-1})],$$

treating the estimated weight function $\widehat{w}^t_h$ as a reward. The `PolicyOptimization` subroutine makes use of exploratory data collected using preceding policy covers $p_1, \ldots, p_{h-1}$, and hence does not have to perform systematic exploration. Indeed, we show that any hybrid offline/online method (that is, any online method that requires access to an exploratory policy) that satisfies a certain "local" policy optimization guarantee is sufficient, with `PSDP` (Algorithm 6; (Bagnell et al., 2003)) and Natural Policy Gradient (Agarwal et al., 2021) being natural choices; for our analysis, we make use of `PSDP`.

[15] Note that while the results in Section 4 are presented for the setting in which the underlying MDP M is "known" to the learner (planning), the algorithmic template in Algorithm 1 can be applied even when M is unknown, as long as the policy optimization step in Line 4 can be implemented in a sample-efficient fashion. This observation is the basis for the algorithms we present in this section.

Algorithm 3 Coverage-Driven Exploration via Weight Function Estimation (`CODEX.W`)

parameters:

Weight function class $\mathcal{W} = \mathcal{W}_{1:H}$.

Policy opt. subroutine $\texttt{PolicyOptimization}_h(r_{1:h}, p_{1:h}, \epsilon, \delta)$. // Optimizes reward r_h using policy covers $p_{1:h}$.

Approximation parameter $\varepsilon \in (0, 1/2)$, failure probability $\delta \in (0, e^{-1})$.

Set $T = \frac{1}{\varepsilon}$ and $p^1 = \pi_{\mathrm{unif}}$.

Set $\epsilon_w = c \cdot (C^{M^\star}_{\mathsf{push}}/|\mathcal{A}|)^{1/2}\varepsilon^{1/2}$, $\epsilon_{\mathsf{opt}} = c' \cdot \varepsilon^2$, and $\delta_w = \delta_{\mathsf{opt}} = \delta/(2HT)$, for suff. small constants $c, c' > 0$.

for $h = 2, \cdots, H$ **do**

  for $t = 1, \ldots, T$ **do**

    $\widehat{w}^t_h \leftarrow \texttt{EstimateWeightFunction}_{h,t}(p_{h-1}, \{\pi^{h,i}\}_{i<t}; \epsilon_w, \delta_w, \mathcal{W})$. // Algorithm 4.

    // Estimate for $w^t_h(x_h \mid x_{h-1}, a_{h-1}) = \frac{P^{M^\star}_{h-1}(x_h|x_{h-1},a_{h-1})}{\sum_{i<t} d^{\pi^{h,i}}_h(x_h) + P^{M^\star}_{h-1}(x_h|x_{h-1},a_{h-1})}$.

    Define reward function $r^{h,t}$ via $r^{h,t}_{h-1}(x_h \mid x_{h-1}, a_{h-1}) = \widehat{w}^t_h(x_h \mid x_{h-1}, a_{h-1})$ and $r^{h,t}_{h'} = 0 \ \ \forall h' \neq h-1$.

    // $r^{h,t}_{h-1}$ can be interpreted as a stochastic reward for layer $h-1$.

    $\pi^{h,t} \leftarrow \texttt{PolicyOptimization}_{h-1}(p_{1:h-1}, r^{h,t}; \epsilon_{\mathsf{opt}}, \delta_{\mathsf{opt}})$.

    // Approximately solve $\arg\max_{\pi\in\Pi} \mathbb{E}^{M^\star,\pi}\left[\frac{P^{M^\star}_{h-1}(x_h|x_{h-1},a_{h-1})}{\sum_{i<t} d^{\pi^{h,i}}_h(x_h) + P^{M^\star}_{h-1}(x_h|x_{h-1},a_{h-1})}\right]$.

  Set $p_h = \mathrm{Unif}(\pi^{h,1} \circ_h \pi_{\mathrm{unif}}, \ldots, \pi^{h,T} \circ_h \pi_{\mathrm{unif}})$.

return Policy covers $(p_1, \ldots, p_H)$.

In what follows, we describe the `EstimateWeightFunction` and `PolicyOptimization` subroutines and the corresponding statistical assumptions in more detail.

Weight function estimation and realizability. To perform weight function estimation, we assume access to a weight function class $\mathcal{W} = \mathcal{W}_{1:H}$, with $\mathcal{W}_h \subseteq (\mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}_+)$ that is capable of representing the weight function w^t_h in Eq. (26). While we can directly assume that the weight function in Eq. (26) is realized by $\mathcal{W}$ (cf. Assumption K.1), it turns out (cf. Proposition K.1) that the following weaker form of weight function realizability is sufficient.

Assumption E.1 (Weight function realizability)**.** *For all $h \geq 2$ and all $\pi \in \Pi_{\mathsf{ns}}$*

$$w^\pi_h(x' \mid x, a) := \frac{P^{M^\star}_{h-1}(x' \mid x, a)}{d^{M^\star,\pi}_h(x')} \in \mathcal{W}_h.$$

Assumption E.1 is new to the best of our knowledge, and is naturally suited to the pushforward coverability objective. While this assumption involves the forward transition probability, it is weaker than model-based realizability because it only requires modeling the *relative* transition probability, as the following example shows.

Example E.1. For a Block MDP (Du et al., 2019; Misra et al., 2020; Zhang et al., 2022; Mhammedi et al., 2023b) with latent state space $\mathcal{S}$ and decoder class $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$, we can satisfy Assumption E.1 with $\log|\mathcal{W}| \leq \widetilde{O}(|\mathcal{S}|^2|\mathcal{A}| + \log|\Phi|)$,[16] yet any realizable model class must have $\log|\mathcal{M}| = \Omega(|\mathcal{X}|) \gg |\mathcal{S}|$ in general. ◁

See Amortila et al. (2024) for further discussion around weight function realizability in online RL.

Our weight function estimation subroutine, `EstimateWeightFunction`, is given in Algorithm 4. To motivate the algorithm, consider the following abstract setting. Let $\mathcal{Z}$ be a set. We receive samples $\mathcal{D}_1 = \{z^1_\mu, \ldots, z^n_\mu\} \in \mathcal{Z}$ and $\mathcal{D}_2 = \{z^1_\nu, \ldots, z^n_\nu\} \in \mathcal{Z}$, where $z^t_\mu \sim \mu^t \in \Delta(\mathcal{Z})$ and $z^t_\nu \sim \nu^t \in \Delta(\mathcal{Z})$. The distributions μ^t and ν^t can be chosen in an adaptive fashion based on $z^1_\mu, z^1_\nu, \ldots, z^{t-1}_\mu, z^{t-1}_\nu$. We define $\mu = \frac{1}{n}\sum_{t=1}^n \mu^t$ and $\nu = \frac{1}{n}\sum_{t=1}^n \nu^t$, and our goal is to estimate the density ratio $w^\star(z) := \frac{\mu(z)}{\nu(z)}$. Given *realizable* weight function class $\mathcal{W}$ with $w^\star \in \mathcal{W}$, Nguyen et al. (2010) (see also

[16] Formally this requires a standard covering argument; we omit the details.

Algorithm 4 Weight Function Estimation ($\texttt{EstimateWeightFunction}_{h,t}(p_{h-1}, \{\pi^i\}_{i<t}; \epsilon, \delta, \mathcal{W})$)

parameters:
Layer $h \geq 2$, iteration $t \in \mathbb{N}$.
Distribution $p_{h-1} \in \Delta(\Pi_{\mathsf{rns}})$, policies $\pi^1, \ldots, \pi^{t-1} \in \Pi_{\mathsf{ns}}$.
Error tolerance $\epsilon \in (0,1)$, failure probability $\delta \in (0,1)$.
Weight function class $\mathcal{W} = \mathcal{W}_{1:H}$ with $\mathcal{W}_h \subseteq (\mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0,1])$.
Let $n = n_{\mathsf{weight}}(\epsilon, \delta) := \frac{40 \log(|\mathcal{W}|\delta^{-1})}{\epsilon^2}$.
Let $q := \frac{1}{2} p_{h-1} + \frac{1}{2(t-1)} \sum_{i<t} \pi^i \circ_{h-1} \pi_{\mathrm{unif}}$ if $t \geq 1$ and $q := p_{h-1}$ otherwise.
Let $\mathcal{D}_1 = \mathcal{D}_2 = \varnothing$.
For each $j \in [n]$, draw $\pi \sim q$ and sample $(x^j_{h-1}, a^j_{h-1}, x^j_h) \sim \pi$. Add $(x^j_{h-1}, a^j_{h-1}, x^j_h)$ to both $\mathcal{D}_1$ and $\mathcal{D}_2$.
for $i < t$ **do**
  Draw n samples $\left\{(x^j_{h-1}, a^j_{h-1}, x^j_h)\right\}_{j \in [n]}$ independently by drawing $\pi \sim q$ and $(x^j_{h-1}, a^j_{h-1}, x^j_h) \sim \pi$.
  Draw n samples $\left\{(\tilde{x}^j_{h-1}, \tilde{a}^j_{h-1}, \tilde{x}^j_h)\right\}_{j \in [n]}$ by sampling $(\tilde{x}^j_{h-1}, \tilde{a}^j_{h-1}, \tilde{x}^j_h) \sim \pi^i$.
  Add $\left\{(x^j_{h-1}, a^j_{h-1}, x^j_h)\right\}_{j \in [n]}$ to $\mathcal{D}_1$ and add $\left\{(x^j_{h-1}, a^j_{h-1}, \tilde{x}^j_h)\right\}_{j \in [n]}$ to $\mathcal{D}_2$.
Set $\widehat{w} := \arg\max_{w \in \mathcal{W}_h} \widehat{\mathbb{E}}_{\mathcal{D}_1}[\log(w(x_h \mid x_{h-1}, a_{h-1}))] - t \cdot \widehat{\mathbb{E}}_{\mathcal{D}_2}[w(x_h \mid x_{h-1}, a_{h-1})]$. // See Eq. (54).
return $\widehat{w}$.

Katdare et al. (2023)) propose the estimator

$$\widehat{w} := \arg\max_{w \in \mathcal{W}} \widehat{\mathbb{E}}_{\mathcal{D}_1}[\log(w)] - \widehat{\mathbb{E}}_{\mathcal{D}_2}[w], \tag{27}$$

where $\widehat{\mathbb{E}}_{\mathcal{D}}[\cdot]$ denotes the empirical expectation with respect to a dataset $\mathcal{D}$. We show (Theorem K.2 in Appendix K) that this estimator ensures that

$$\mathbb{E}_{z \sim \nu}\left[\left(\sqrt{\widehat{w}(z)} - \sqrt{w^\star(z)}\right)^2\right] \lesssim O(\|w^\star\|_\infty) \cdot \frac{\log(|\mathcal{W}|\delta^{-1})}{n} \tag{28}$$

with high probability. Algorithm 4 simply applies this technique to estimate the weight function in Eq. (26); to ensure realizability of Eq. (26) under Assumption E.1, we apply the method with an expanded weight function class (defined in Proposition K.1); see Appendix K.1 for details.

Remark E.1 (Comparison to contrastive learning (Misra et al., 2020))**.** We remark our weight function estimation subroutine (Algorithm 4) can be viewed as a form of contrastive learning, with the target (population) weight function in Eq. (26) bearing strong similarity to the target (Bayes-optimal) function from the regression problem used in the HOMER algorithm (cf. Line 11 of Algorithm 3 and Lemma 9 of Misra et al.). We also note that both CODEX.W and HOMER find policy covers inductively via policy optimization on exploratory reward functions, and thus we can view CODEX.W as a natural generalization of the HOMER algorithm to settings beyond the Block MDP. Additionally, our analysis improves dependencies present in HOMER's sample complexity (notably, the dependence on the minimal visitation probability).

Policy optimization subroutine. The policy optimization subroutine, PolicyOptimization, is a hyperparameter to the algorithm, and can be any subroutine that approximately solves the policy optimization problem

$$\pi^{h,t} \approx \arg\max_{\pi \in \Pi} \mathbb{E}^{M^\star, \pi}[\widehat{w}^t_h(x_h \mid x_{h-1}, a_{h-1})],$$

which treats the estimated weight function $\widehat{w}^t_h$ as a reward for layer $h-1$. PolicyOptimization does not need to perform global policy optimization—instead, we only require a certain form of "local" guarantee with respect to the approximate policy covers $p_1, \ldots, p_{h-1}$; see Assumption K.2 for details.

For concreteness, we make use PSDP (Bagnell et al., 2003), described in Appendix K.4, as the PolicyOptimization subroutine. PSDP optimizes a given reward function r_{h-1} by collecting exploratory data with $p_1, \ldots, p_{h-1}$ and applying

approximate dynamic programming with a user-specified value function class $\mathcal{Q}$. We make the following value-based realizability assumption, which asserts that $\mathcal{Q}$ is expressive enough to allow PSDP to optimize any weight function in the class $\mathcal{W}$.

Assumption E.2. *We have access to a value function class $\mathcal{Q} = \mathcal{Q}_{1:H}$ with $\mathcal{Q}_h \subseteq (\mathcal{X} \times \mathcal{A} \to [0,1])$ such that for all $h \in [H]$, $w \in \mathcal{W}_h$, and $\pi \in \Pi_{\mathsf{ns}}$, we have $Q_\ell^{M^\star,\pi}(\cdot,\cdot;w) \in \mathcal{Q}_\ell$ for all $\ell \le h-1$, where*

$$Q_\ell^{M^\star,\pi}(x,a;w) = \mathbb{E}^{M^\star,\pi}[w_h(x_h \mid x_{h-1},a_{h-1}) \mid x_\ell = x, a_\ell = a]$$

is the Q-function for π under the (stochastic) reward function defined via $r_{h-1} = w(x_h \mid x_{h-1},a_{h-1})$ and $r_{h'} = 0$ for $h' \neq h-1$.

E.2. Main Result

The main guarantee for Algorithm 3 applied with PSDP is as follows.

Theorem E.1 (Main result for Algorithm 3 with PSDP)**.** *Let $\varepsilon \in (0,1/2)$ and $\delta \in (0,e^{-1})$ be given, and suppose that we have a weight function class $\mathcal{W}$ and value function class $\mathcal{Q}$ such that Assumptions E.1 and E.2 are satisfied. Then with PSDP (Algorithm 6) as a policy optimization subroutine, and using an expanded weight function class defined in Proposition K.1, Algorithm 3 produces policy covers $p_1,\dots,p_H \in \Delta(\Pi_{\mathsf{rns}})$ such that with probability at least $1-\delta$, for all $h \in [H]$,*

$$\Psi^{M^\star}_{\mathsf{push};h,\varepsilon}(p_h) \le 170H\log(\varepsilon^{-1}) \cdot C^{M^\star}_{\mathsf{push}},$$

and does so using at most

$$N \le \widetilde{O}\left(\frac{H|\mathcal{A}|\log(|\mathcal{W}|\delta^{-1})}{\varepsilon^4} + \frac{H^4|\mathcal{A}|\log(|\mathcal{Q}|\delta^{-1})}{\varepsilon^5}\right)$$

episodes.

Theorem E.1 is a special case of a more general result (Theorem K.1) which allows for general policy optimization algorithms that satisfy a certain "local optimality" guarantee (Assumption K.2); we obtain the result by verifying that PSDP satisfies this condition. Let us discuss some key features.

- *Sample efficiency.* The sample complexity in Theorem E.1 scales as $\mathrm{poly}(H,|\mathcal{A}|,\log|\mathcal{W}|,\log|\mathcal{Q}|,\varepsilon^{-1})$, and hence is efficient for large state spaces; we expect that the precise polynomial factors can be tightened, as can the approximation ratio in the objective value, though we leave this for future work. As with our results in preceding sections, the resulting policy covers $p_1,\dots,p_H$ can be directly used for downstream policy optimization.
- *Computational efficiency.* With PSDP as the `PolicyOptimization` subroutine, Algorithm 3 is computationally efficient whenever i) the weight function estimation objective in Line 10 of Algorithm 4 can be solved efficiently over $\mathcal{W}$, and ii) square loss regression over the class $\mathcal{Q}$ can be solved efficiently. We consider these to be fairly mild assumptions.
- *Practicality.* In practice, we expect that the subroutine `PolicyOptimization` can be implemented using off-the-shelf deep RL methods (e.g., PPO or SAC), which are known to perform well given access to exploratory data. In this sense, Algorithm 3 can be viewed as a new approach to equipping existing deep RL methods with exploration, with the weight function-based rewards in Line 7 acting as exploration bonuses.

As a concrete example, we can instantiate Theorem E.1 for Block MDPs to derive end-to-end guarantees under standard assumptions.

Example E.2 (Sample complexity for Block MDP)**.** For a Block MDP (Du et al., 2019; Misra et al., 2020; Zhang et al., 2022; Mhammedi et al., 2023b) with latent state space $\mathcal{S}$ and decoder class $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$, we can satisfy Assumption E.1 with $\log|\mathcal{W}| \le \widetilde{O}(|\mathcal{S}|^2|\mathcal{A}| + \log|\Phi|)$ and $\log|\mathcal{Q}| \le \widetilde{O}(|\mathcal{S}||\mathcal{A}| + \log|\Phi|)$, so Theorem E.1 gives sample complexity $N \le \widetilde{O}\left(\frac{H^4|\mathcal{S}|^2|\mathcal{A}|^2\log(|\Phi|\delta^{-1})}{\varepsilon^5}\right)$. ◁

We view the results in this section as a proof of concept, showing that the planning methods derived in Section 4 for iteratively optimizing L_1-Coverage can be implemented in a data-driven fashion. We leave sample-efficient counterparts for other relaxations for future work.

F. L_1-Coverage: Structural Properties

In this section we draw connections between L_1-Coverability, the structural parameter induced by L_1-Coverage, and other structural parameters and objectives, focusing on (i) a non-admissible variant of L_1-Coverability (Appendix F.1), and (ii) feature coverage (Appendix F.2). We also show that alternative exploration objectives *do not* induce meaningful structural parameters in the same fashion as L_1-Coverage (Appendix F.3).

F.1. Connection to Non-Admissible L_1-Coverage

Inspired by Eq. (5), we can also define a non-admissible counterpart to L_1-Coverability as follows.

$$C^M_{\mathsf{avg};h} = \inf_{\mu\in\Delta(\mathcal{X}\times\mathcal{A})} \sup_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{d^{M,\pi}_h(x_h,a_h)}{\mu(x_h,a_h)}\right], \tag{29}$$

and $C^M_{\mathsf{avg}} = \max_{h\in[H]} C^M_{\mathsf{avg};h}$. This quantity was used to provide sample complexity bounds for online reinforcement learning by Liu et al. (2023a) (generalizing the results of Xie et al. (2023)), and the associated concentrability coefficient $C^M_{\mathsf{avg};h}(\mu) = \sup_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{d^{M,\pi}_h(x_h,a_h)}{\mu(x_h,a_h)}\right]$ is widely used in offline reinforcement learning (Farahmand et al., 2010; Xie & Jiang, 2020). The following result, which uses the minimax theorem in a similar fashion to Proposition 3.2, shows that it is possible to bound L_1-Coverability by this quantity in spite of non-admissibility, albeit with some loss in rate.

Proposition F.1. *For all $\varepsilon > 0$, it holds that $\mathsf{Cov}^M_{h,\varepsilon} \leq 1 + 2\sqrt{\frac{C^M_{\mathsf{avg};h}}{\varepsilon}}$.*

Note that while the bound in Proposition F.1 grows as $\sqrt{\frac{1}{\varepsilon}}$, it still leads to non-trivial sample complexity bounds through our main results (which give meaningful guarantees whenever $\mathsf{Cov}^M_{h,\varepsilon}$ grows sublinearly with ε^{-1}), though the resulting rates are worse than in the case where $\mathsf{Cov}^M_{h,\varepsilon}$ is bounded by an absolute constant.

F.2. Connection to Feature Coverage

Another well-studied notion of coverage from offline reinforcement learning is *feature coverage* (Jin et al., 2021b; Zanette et al., 2021; Wagenmaker & Pacchiano, 2023). Consider the Low-Rank MDP framework (Rendle et al., 2010; Yao et al., 2014; Agarwal et al., 2020; Modi et al., 2024; Zhang et al., 2022; Mhammedi et al., 2023a), in which the transition distribution is assumed to factorize as

$$P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1}) = \langle \phi_{h-1}(x_{h-1}, a_{h-1}), \psi_h(x_h)\rangle, \tag{30}$$

where $\phi_{h-1}(x,a), \psi_h(x) \in \mathbb{R}^d$ are (potentially unknown) feature maps. For offline reinforcement learning in Low-Rank MDPs (Jin et al., 2021b; Zanette et al., 2021; Wagenmaker & Pacchiano, 2023), feature coverage for a distribution $\mu \in \Delta(\mathcal{X}\times\mathcal{A})$ is given by $C^M_{\phi;h}(\mu) = \sup_{\pi\in\Pi} \|\mathbb{E}^{M,\pi}[\phi_h(x_h,a_h)]\|^2_{\Sigma_\mu^{-1}}$. We define the associated feature coverability coefficient by

$$C^M_{\phi;h} = \inf_{\mu\in\Delta(\mathcal{X}\times\mathcal{A})} \sup_{\pi\in\Pi} \|\mathbb{E}^{M,\pi}[\phi_h(x_h,a_h)]\|^2_{\Sigma_\mu^{-1}}, \tag{31}$$

where $\Sigma_\mu := \mathbb{E}_{(x,a)\sim\mu}\left[\phi_h(x,a)\phi_h(x,a)^\top\right]$, and define $C^M_\phi = \max_{h\in[H]} C^M_{\phi;h}$. Note that one always has $C^M_\phi \leq d$, as a consequence of the existence of G-optimal designs (Kiefer & Wolfowitz, 1960; Huang et al., 2023). The following result shows that L_1-Coverability is always bounded by feature coverability.

Proposition F.2. *Suppose the MDP M obeys the low-rank structure in Eq. (30). Then for all $h \in [H]$, we have $C^M_{\mathsf{avg};h} \leq |\mathcal{A}| \cdot C^M_{\phi;h-1}$, and consequently $\mathsf{Cov}^M_{h,\varepsilon} \leq 1 + 2\sqrt{\frac{|\mathcal{A}|\cdot C^M_{\phi;h-1}}{\varepsilon}}$.*

We remark that if we restrict the distribution $\mu \in \Delta(\mathcal{X}\times\mathcal{A})$ in Eq. (31) to be realized as a mixture of occupancies, Proposition F.2 can be strengthened to give $\mathsf{Cov}^M_{h,\varepsilon} \leq O\big(|\mathcal{A}| \cdot C^M_{\phi;h-1}\big)$. Note that through our main results, Proposition F.2, gives guarantees that scale with $|\mathcal{A}|$. This is necessary in the setting where the feature map ϕ_h is unknown, which is covered by our results, but is suboptimal in the setting where ϕ is known.

F.3. Insufficiency of Alternative Notions of Coverage

L_1-Coverage measures coverage of the mixture policy $p \in \Delta(\Pi_{\mathsf{rns}})$ with respect to the $L_1(d_h^{M,\pi})$-norm. It is also reasonable to consider variants of the objective based on L_q-norms for $q > 1$, which provide stronger coverage, but may have larger optimal value. In particular, a natural L_∞-type analogue of Eq. (2) is given by

$$\Psi^M_{\infty;h,\varepsilon}(p) = \sup_{\pi\in\Pi} \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \left\{ \frac{d_h^{M,\pi}(x,a)}{d_h^{M,p}(x,a) + \varepsilon \cdot d_h^{M,\pi}(x,a)} \right\}. \tag{32}$$

is identical to the L_∞-coverability coefficient C^M_∞ (5) studied in Xie et al. (2023), except that we restrict the data distribution $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$ to be admissible, i.e. realized by a mixture policy $p \in \Delta(\Pi_{\mathsf{rns}})$ (we also incorporate the term $\varepsilon \cdot d_h^{M,\pi}(x,a)$ in the denominator to ensure the ratio is well-defined). The following lemma shows, perhaps surprisingly, that in stark contrast to L_1-Coverage, it is not possible to bound the optimal value of the admissible L_∞-Coverage objective in Eq. (32) in terms of the non-admissible coverability coefficient C^M_∞.

Proposition F.3. *There exists an MDP M and policy class $\Pi \subset \Pi_{\mathsf{rns}}$ with horizon $H = 1$ such that $C^M_{\infty;h} \le 2$ (and hence $\mathsf{Cov}^M_{h,\varepsilon} \le 2$ as well), yet for all $\varepsilon > 0$,*

$$\inf_{p\in\Delta(\Pi)} \Psi^M_{\infty;h,\varepsilon}(p) \ge \frac{1}{\varepsilon}, \tag{33}$$

and in particular $\inf_{p\in\Delta(\Pi)} \Psi^M_{\infty;h,0}(p) = \infty$.

Note that one trivially has $\inf_{p\in\Delta(\Pi)} \Psi^M_{\infty;h,\varepsilon}(p) \le \frac{1}{\varepsilon}$, and hence Eq. (33) shows that the optimal value is vacuously large for the MDP in this example. In contrast, we have $\mathsf{Cov}^M_{h,\varepsilon} \le 2$ even when $\varepsilon = 0$. More generally, one can show that similar failure modes hold for admissible L_q-Coverage for any $q > 1$.

Part II

Proofs

G. Proofs from Section 3

Proposition 3.1 (Change of measure for L_1-Coverage)**.** *For any distribution $p \in \Delta(\Pi_{\mathsf{rns}})$, we have that for all functions $g : \mathcal{X} \times \mathcal{A} \to [0, B]$, all $\pi \in \Pi$, and all $\varepsilon > 0$,*[17]

$$\begin{aligned} &\mathbb{E}^{M,\pi}[g(x_h, a_h)] \\ &\leq 2\sqrt{\Psi^M_{h,\varepsilon}(p) \cdot \mathbb{E}^{M,p}[g^2(x_h, a_h)]} + \Psi^M_{h,\varepsilon}(p) \cdot (\varepsilon B). \end{aligned} \tag{3}$$

Proof of Proposition 3.1. Let $p \in \Delta(\Pi_{\mathsf{rns}})$ and $g : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ be given. We first prove the following, more general inequality:

$$\mathbb{E}^{M,\pi}[g(x_h, a_h)] \leq \sqrt{\Psi^M_{h,\varepsilon}(p) \cdot (\mathbb{E}^{M,p}[g^2(x_h, a_h)] + \varepsilon \cdot \mathbb{E}^{M,\pi}[g^2(x_h, a_h)])}. \tag{34}$$

Using Cauchy-Schwarz, we have

$$\begin{aligned} \mathbb{E}^{M,\pi}[g(x_h, a_h)] &= \sum_{x \in \mathcal{X}, a \in \mathcal{A}} d^{M,\pi}_h(x, a) g(x, a) \\ &= \sum_{x \in \mathcal{X}, a \in \mathcal{A}} d^{M,\pi}_h(x, a) \cdot \frac{(d^{pM,p}_h(x, a) + \varepsilon \cdot d^{M,\pi}_h(x, a))^{1/2}}{(d^{M,p}_h(x, a) + \varepsilon \cdot d^{M,\pi}_h(x, a))^{1/2}} \cdot g(x, a) \\ &\leq \left(\sum_{x \in \mathcal{X}, a \in \mathcal{A}} \frac{(d^{M,\pi}_h(x, a))^2}{d^{M,p}_h(x, a) + \varepsilon \cdot d^{M,\pi}_h(x, a)} \right)^{1/2} \left(\sum_{x \in \mathcal{X}, a \in \mathcal{A}} (d^{M,p}_h(x, a) + \varepsilon \cdot d^{M,\pi}_h(x, a)) g^2(x, a) \right)^{1/2} \\ &= \sqrt{\Psi^M_{h,\varepsilon}(p) \cdot (\mathbb{E}^{M,p}[g^2(x_h, a_h)] + \varepsilon \cdot \mathbb{E}^{M,\pi}[g^2(x_h, a_h)])}. \end{aligned}$$

This establishes Eq. (34). To prove Eq. (3), we first bound

$$\begin{aligned} \sqrt{\Psi^M_{h,\varepsilon}(p) \cdot (\mathbb{E}^{M,p}[g^2(x_h, a_h)] + \varepsilon \cdot \mathbb{E}^{M,\pi}[g^2(x_h, a_h)])} &\leq \sqrt{\Psi^M_{h,\varepsilon}(p) \cdot \mathbb{E}^{M,p}[g^2(x_h, a_h)]} \\ &\quad + \sqrt{\Psi^M_{h,\varepsilon}(p) \cdot \varepsilon \cdot \mathbb{E}^{M,\pi}[g^2(x_h, a_h)]}. \end{aligned}$$

Next, we note that if $g \in [0, B]$, we can use AM-GM to bound

$$\sqrt{\Psi^M_{h,\varepsilon}(p) \cdot \varepsilon \cdot \mathbb{E}^{M,\pi}[g^2(x_h, a_h)]} \leq \sqrt{\Psi^M_{h,\varepsilon}(p) \cdot (\varepsilon B) \cdot \mathbb{E}^{M,\pi}[g(x_h, a_h)]} \leq \frac{\Psi^M_{h,\varepsilon}(p) \cdot (\varepsilon B)}{2} + \frac{1}{2} \mathbb{E}^{M,\pi}[g(x_h, a_h)].$$

The result now follows by rearranging. □

Proposition 3.2. *For all $\varepsilon > 0$, we have $\mathsf{Cov}^M_{h,\varepsilon} \leq C^M_{\infty;h}$.*

Proof of Proposition 3.2. Let $\delta > 0$ be given. Using Lemma C.10 and the definition of $C^M_{\infty;h}$, there exists $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$ such that

$$\begin{aligned} \mathsf{Cov}^M_{h,\varepsilon} &\leq \left(1 + \frac{\delta}{\varepsilon}\right) C^M_{\infty;h} \cdot \inf_{p \in \Delta(\Pi)} \sup_{\pi \in \Pi} \mathbb{E}^{M,\pi} \left[\frac{\mu(x_h, a_h)}{d^{M,p}_h(x_h, a_h) + \delta \cdot C^M_{\infty;h} \mu(x_h, a_h)} \right] \\ &= \left(1 + \frac{\delta}{\varepsilon}\right) C^M_{\infty;h} \cdot \inf_{p \in \Delta(\Pi)} \sup_{q \in \Delta(\Pi)} \mathbb{E}_{\pi \sim q} \mathbb{E}^{M,\pi} \left[\frac{\mu(x_h, a_h)}{d^{M,p}_h(x_h, a_h) + \delta \cdot C^M_{\infty;h} \mu(x_h, a_h)} \right]. \end{aligned}$$

[17] This result is meaningful in the parameter regime where $\Psi^M_{h,\varepsilon}(p) < 1/\varepsilon$. We refer to this regime as *non-trivial*, as $\Psi^M_{h,\varepsilon}(p) \leq 1/\varepsilon$ holds vacuously for all p.

Observe that the function

$$(p,q) \mapsto \mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,p}(x_h,a_h)+\delta\cdot C^M_{\infty;h}\mu(x_h,a_h)}\right]$$

is convex-concave. In addition, it is straightforward to see that the function is jointly Lipschitz with respect to total variation distance whenever $\varepsilon,\delta>0$. Hence, using the minimax theorem (Lemma C.1), we have that

$$\begin{aligned}
&\inf_{p\in\Delta(\Pi)}\sup_{q\in\Delta(\Pi)}\mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,p}(x_h,a_h)+\delta\cdot C^M_{\infty;h}\mu(x_h,a_h)}\right]\\
&=\sup_{q\in\Delta(\Pi)}\inf_{p\in\Delta(\Pi)}\mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,p}(x_h,a_h)+\delta\cdot C^M_{\infty;h}\mu(x_h,a_h)}\right]\\
&\le\sup_{q\in\Delta(\Pi)}\mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,q}(x_h,a_h)+\delta\cdot C^M_{\infty;h}\mu(x_h,a_h)}\right]\\
&=\sum_{x\in\mathcal{X},a\in\mathcal{A}}\frac{d_h^{M,q}(x,a)\mu(x,a)}{d_h^{M,q}(x,a)+\delta\cdot C^M_{\infty;h}\mu(x,a)}\le 1.
\end{aligned}$$

To conclude, we take $\delta\to 0$. □

H. Proofs and Additional Details from Section 4

H.1. Omitted Algorithms

Algorithm 5 Approximate Policy Cover Computation via Pushforward Coverability Relaxation

1: **input**: Layer $h\in[H]$, precision parameter , $\varepsilon\in[0,1]$, optimization tolerance $\varepsilon_{\mathrm{opt}}>0$.
2: Set $T=\frac{1}{\varepsilon}$.
3: **for** $t=1,2,\cdots,T$ **do**
4: Compute $\pi^t\in\Pi$ such that

$$\mathbb{E}^{M,\pi^t}\left[\frac{P^M_{h-1}(x_h\mid x_{h-1},a_{h-1})}{\sum_{i<t}d_h^{M,\pi^i}(x_h)+P^M_{h-1}(x_h\mid x_{h-1},a_{h-1})}\right]\ge\sup_{\pi\in\Pi}\mathbb{E}^{M,\pi}\left[\frac{P^M_{h-1}(x_h\mid x_{h-1},a_{h-1})}{\sum_{i<t}d_h^{M,\pi^i}(x_h)+P^M_{h-1}(x_h\mid x_{h-1},a_{h-1})}\right]-\varepsilon_{\mathrm{opt}}.$$

5: Return $p=\mathrm{Unif}(\pi^1,\dots,\pi^T)$.

H.2. Examples for Algorithm 1

As discussed in Section 4.1, the L_∞-coverability relaxation (6) used by Algorithm 1 can be optimized efficiently whenever a (non-admissible) state-action distribution $\mu\in\Delta(\mathcal{X}\times\mathcal{A})$ with low L_∞-concentrability $C^M_{\infty;h}$ can be computed efficiently for the MDP M. Examples of MDP classes the admit efficiently computable distributions with low concentrability include:

- When M is a tabular MDP, the distribution $\mu(x,a)=\frac{1}{|\mathcal{X}||\mathcal{A}|}$ (which clearly admits a closed form representation) achieves $C^M_{\infty;h}(\mu)\le|\mathcal{X}||\mathcal{A}|$.
- For a Block MDPs (Du et al., 2019; Misra et al., 2020; Zhang et al., 2022; Mhammedi et al., 2023b) with latent state space $\mathcal{S}$, emission distribution $q:\mathcal{S}\to\Delta(\mathcal{X})$, and decoder $\phi^\star:\mathcal{X}\to\mathcal{S}$, the distribution $\mu(x,a):=q(x\mid\phi^\star(x))\cdot\frac{1}{|\mathcal{S}||\mathcal{A}|}$ achieves $C^M_{\infty;h}(\mu)\le|\mathcal{S}||\mathcal{A}|$ (Xie et al., 2023). Again, this distribution admits a closed form representation when M is explicitly specified.
- For low-rank MDPs with the structure in Eq. (30), the distribution given by $\mu(x,a)=\frac{\|\psi_h(x)\|_2}{\int\|\psi_h(x')\|_2dx'}\cdot\frac{1}{|\mathcal{A}|}$ achieves $C^M_{\infty;h}(\mu)\le B|\mathcal{A}|$ under the standard normalization assumption that $\int\|\psi_h(x')\|_2dx'\le B$ and $\|\phi(x,a)\|_2\le 1$ for some

(typically dimension-dependent) constant $B > 0$ (Mhammedi et al., 2023a; Golowich et al., 2023). Alternatively we can compute a set of policies $\pi^1, \ldots, \pi^d$ that form a barycentric spanner for the set $\{\mathbb{E}^{M,\pi}[\phi(x_h, a_h)]\}_{\pi\in\Pi}$ and choose $\mu(x, a) = \frac{1}{d|\mathcal{A}|}\sum_{i=1}^{d} d_h^{M,\pi^i}(x)$, which achieves $C^M_{\infty;h}(\mu) \le d|\mathcal{A}|$ (Huang et al., 2023).

These examples highlight that for many settings of interest, computing a covering distribution $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$ when the model is known is significantly simpler than computing an explicit policy cover $p \in \Delta(\Pi_{\mathsf{rns}})$, showcasing the utility of Algorithm 1.

H.3. Proofs from Section 4.1

Proposition 4.1. *For a distribution μ with $C_\infty \equiv C^M_{\infty;h}(\mu)$, it holds that for all $p \in \Delta(\Pi_{\mathsf{rns}})$,*

$$\Psi^M_{h,\varepsilon}(p) \le 2C_\infty \cdot \Psi^M_{\mu;h,\varepsilon}(p). \tag{7}$$

Furthermore, $\mathsf{Cov}^M_{\mu;h,\varepsilon} \le 1$ for all $\varepsilon > 0$.

Proof of Proposition 4.1. Fix μ and abbreviate $C_\infty \equiv C^M_{\infty;h}(\mu)$. Observe that for any $\pi \in \Pi$ and $p \in \Delta(\Pi_{\mathsf{rns}})$, Lemma C.10 implies that we can bound

$$\begin{aligned}
&\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h, a_h)}{d_h^{M,p}(x_h, a_h) + \varepsilon \cdot d_h^{M,\pi}(x_h, a_h)}\right] \\
&\le \mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h, a_h)}{d_h^{M,p}(x_h, a_h) + \varepsilon \cdot C_\infty\mu(x_h, a_h)}\right] + C_\infty\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h, a_h)}{d_h^{M,p}(x_h, a_h) + \varepsilon \cdot C_\infty\mu(x_h, a_h)}\right] \\
&\le 2C_\infty\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h, a_h)}{d_h^{M,p}(x_h, a_h) + \varepsilon \cdot C_\infty\mu(x_h, a_h)}\right].
\end{aligned}$$

For the claim that $\mathsf{Cov}^M_{\mu;h,\varepsilon} \le 1$, see the proof of Proposition 3.2.

□

Theorem 4.1. *For any $\varepsilon \in [0, 1]$ and $h \in [H]$, given a distribution μ with $C_\infty \equiv C^M_{\infty;h}(\mu)$, whenever $\varepsilon_{\mathrm{opt}} \le \varepsilon \log(2\varepsilon^{-1})$, Algorithm 1 with $T = \varepsilon^{-1}$ produces a distribution $p \in \Delta(\Pi)$ with $|\mathrm{supp}(p)| \le \varepsilon^{-1}$ such that*

$$\Psi^M_{\mu;h,\varepsilon}(p) \le 3\log(2\varepsilon^{-1}), \tag{9}$$

and consequently $\Psi^M_{h,\varepsilon}(p) \le 6C_\infty \log(2\varepsilon^{-1})$.

Proof of Theorem 4.1. Let us abbreviate $\widetilde{d}^t = \sum_{i<t} d^{M,\pi^i}$. Observe that for $T = \frac{1}{\varepsilon}$, we have

$$\Psi^M_{\mu;h,\varepsilon}(p) = \sup_{\pi\in\Pi}\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h, a_h)}{\mathbb{E}_{\pi'\sim p}\left[d_h^{M,\pi'}(x_h, a_h)\right] + \frac{C_\infty}{T}\mu(x_h, a_h)}\right] = T \cdot \sup_{\pi\in\Pi}\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h, a_h)}{\widetilde{d}_h^{T+1}(x_h, a_h) + C_\infty\mu(x_h, a_h)}\right],$$

and hence it suffices to bound the quantity on the right-hand side. Observe that for all $t \in [T]$, we have that

$$\sup_{\pi\in\Pi}\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h, a_h)}{\widetilde{d}_h^{t}(x_h, a_h) + C_\infty\mu(x_h, a_h)}\right] \le \sup_{\pi\in\Pi}\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h, a_h)}{\widetilde{d}_h^{t-1}(x_h, a_h) + C_\infty\mu(x_h, a_h)}\right],$$

and consequently

$$\begin{aligned}
T \cdot \sup_{\pi\in\Pi}\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h, a_h)}{\widetilde{d}_h^{T+1}(x_h, a_h) + C_\infty\mu(x_h, a_h)}\right] &\le \sum_{t=1}^{T}\sup_{\pi\in\Pi}\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h, a_h)}{\widetilde{d}_h^{t}(x_h, a_h) + C_\infty\mu(x_h, a_h)}\right] \\
&\le \sum_{t=1}^{T}\mathbb{E}^{M,\pi^t}\left[\frac{\mu(x_h, a_h)}{\widetilde{d}_h^{t}(x_h, a_h) + C_\infty\mu(x_h, a_h)}\right] + \varepsilon_{\mathrm{opt}}T.
\end{aligned}$$

Finally, we note that

$$\sum_{t=1}^{T} \mathbb{E}^{M,\pi^t}\left[\frac{\mu(x_h,a_h)}{\widetilde{d}_h^t(x_h,a_h)+C_\infty\mu(x_h,a_h)}\right] = \sum_{x\in\mathcal{X},a\in\mathcal{A}}\sum_{t=1}^{T}\mu(x,a)\frac{d^{M,\pi^t}(x,a)}{\widetilde{d}_h^t(x,a)+C_\infty\mu(x,a)}.$$

Since $\sup_{\pi\in\Pi} d_h^{M,\pi}(x,a) \le C_\infty\mu(x,a)$ for all $(x,a)\in\mathcal{X}\times\mathcal{A}$, Lemma C.9 implies that

$$\sum_{x\in\mathcal{X},a\in\mathcal{A}}\sum_{t=1}^{T}\mu(x,a)\frac{d^{M,\pi^t}(x,a)}{\widetilde{d}_h^t(x,a)+C_\infty\mu(x,a)} \le 2\log(2T),$$

allowing us to conclude that

$$\Psi^M_{\mu;h,\varepsilon}(p) \le 2\log(2T)+\varepsilon_{\mathrm{opt}}T.$$

□

H.4. Proofs from Section 4.2

Proposition 4.2. *Fix $h\in[H]$. For any $p\in\Delta(\Pi_{\mathsf{rns}})$, if we define $p'\in\Delta(\Pi_{\mathsf{rns}})$ as the distribution induced by sampling $\pi\sim p$ and executing $\pi\circ_h\pi_{\mathrm{unif}}$, we have that for all $\varepsilon>0$,*

$$\Psi^M_{h,\varepsilon}(p') \le |\mathcal{A}|\cdot\Psi^M_{\mathsf{push};h,\varepsilon}(p). \tag{11}$$

Furthermore, $\mathsf{Cov}^M_{\mathsf{push};h,\varepsilon} \le C^M_{\mathsf{push};h}$ for all $\varepsilon>0$.

Proof of Proposition 4.2. We first note that

$$\Psi^M_{h,\varepsilon}(p') \le |\mathcal{A}|\cdot\sup_{\pi\in\Pi}\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h)}{d_h^{M,p}(x_h)+\varepsilon\cdot d_h^{M,\pi}(x_h)}\right].$$

Next, we write

$$\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h)}{d_h^{M,p}(x_h)+\varepsilon\cdot d_h^{M,\pi}(x_h)}\right] = \sum_{x\in\mathcal{X}}\frac{(d_h^{M,\pi}(x))^2}{d_h^{M,p}(x)+\varepsilon\cdot d_h^{M,\pi}(x)}.$$

We now state and prove a basic technical lemma.

Lemma H.1. *For all $\varepsilon,\delta>0$, the function $f(x)=\frac{x^2}{\delta+\varepsilon x}$ is convex over $\mathbb{R}_+$.*

Proof of Lemma H.1. This follows by verifying through direct calculation that

$$f'(x)=\varepsilon\cdot\frac{x^2}{(\delta+\varepsilon x)^2},\quad\text{and}\quad f''(x)=4\varepsilon\delta\cdot\frac{x}{(\delta+\varepsilon x)^3}\ge 0.$$

□

By Lemma H.1, the function

$$d\mapsto\frac{(d)^2}{d_h^{M,p}(x)+\varepsilon\cdot d}$$

is convex for all x. Hence, writing $d_h^{M,\pi}(x)=\mathbb{E}^{M,\pi}\left[P^M_{h-1}(x\mid x_{h-1},a_{h-1})\right]$, Jensen's inequality implies that for all x,

$$\frac{(d_h^{M,\pi}(x))^2}{d_h^{M,p}(x)+\varepsilon\cdot d_h^{M,\pi}(x)} \le \mathbb{E}^{M,\pi}\left[\frac{(P^M_{h-1}(x\mid x_{h-1},a_{h-1}))^2}{d_h^{M,p}(x)+\varepsilon\cdot P^M_{h-1}(x\mid x_{h-1},a_{h-1})}\right].$$

We conclude that

$$\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h)}{d_h^{M,p}(x_h)+\varepsilon\cdot d_h^{M,\pi}(x_h)}\right] \le \mathbb{E}^{M,\pi}\left[\sum_{x\in\mathcal{X}}\frac{(P_{h-1}^M(x\mid x_{h-1},a_{h-1}))^2}{d_h^{M,p}(x)+\varepsilon\cdot P_{h-1}^M(x\mid x_{h-1},a_{h-1})}\right]$$
$$= \mathbb{E}^{M,\pi}\left[\frac{P_{h-1}^M(x_h\mid x_{h-1},a_{h-1})}{d_h^{M,p}(x)+\varepsilon\cdot P_{h-1}^M(x_h\mid x_{h-1},a_{h-1})}\right] \le \Psi^M_{\mathsf{push};h,\varepsilon}(p).$$

We now prove the bound on $\mathsf{Cov}^M_{\mathsf{push};h,\varepsilon}$. Let $\delta>0$ be given. Using the definition of $C^M_{\mathsf{push};h}$ and the same argument as Lemma C.10, there exists $\mu\in\Delta(\mathcal{X})$ such that

$$\mathsf{Cov}^M_{\mathsf{push};h,\varepsilon} \le \left(1+\frac{\delta}{\varepsilon}\right)C^M_{\mathsf{push};h}\cdot\inf_{p\in\Delta(\Pi)}\sup_{\pi\in\Pi}\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h)}{d_h^{M,p}(x_h)+\delta\cdot C^M_{\mathsf{push};h}\mu(x_h)}\right]$$
$$= \left(1+\frac{\delta}{\varepsilon}\right)C^M_{\mathsf{push};h}\cdot\inf_{p\in\Delta(\Pi)}\sup_{q\in\Delta(\Pi)}\mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h)}{d_h^{M,p}(x_h)+\delta\cdot C^M_{\mathsf{push};h}\mu(x_h)}\right].$$

Observe that the function

$$(p,q)\mapsto \mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h)}{d_h^{M,p}(x_h)+\delta\cdot C^M_{\mathsf{push};h}\mu(x_h)}\right]$$

is convex-concave. In addition, it is straightforward to see that the function is jointly Lipschitz with respect to total variation distance whenever $\varepsilon,\delta>0$. Hence, using the minimax theorem (Lemma C.1), we have that

$$\inf_{p\in\Delta(\Pi)}\sup_{q\in\Delta(\Pi)}\mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h)}{d_h^{M,p}(x_h)+\delta\cdot C^M_{\mathsf{push};h}\mu(x_h)}\right]$$
$$= \sup_{q\in\Delta(\Pi)}\inf_{p\in\Delta(\Pi)}\mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h)}{d_h^{M,p}(x_h)+\delta\cdot C^M_{\mathsf{push};h}\mu(x_h)}\right]$$
$$\le \sup_{q\in\Delta(\Pi)}\mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h)}{d_h^{M,q}(x_h)+\delta\cdot C^M_{\mathsf{push};h}\mu(x_h)}\right]$$
$$= \sum_{x\in\mathcal{X}}\frac{d_h^{M,q}(x)\mu(x)}{d_h^{M,q}(x)+\delta\cdot C^M_{\mathsf{push};h}\mu(x)} \le 1.$$

To conclude, we take $\delta\to 0$.

□

Theorem 4.2. *For any $\varepsilon\in[0,1]$ and $h\in[H]$, whenever $\varepsilon_{\mathrm{opt}}\le C^M_{\mathsf{push};h}\cdot\varepsilon\log(2\varepsilon^{-1})$, Algorithm 5 produces a distribution $p\in\Delta(\Pi)$ with $|\mathrm{supp}(p)|\le\varepsilon^{-1}$ such that*

$$\Psi^M_{\mathsf{push};h,\varepsilon}(p)\le 5C^M_{\mathsf{push};h}\log(2\varepsilon^{-1}). \tag{12}$$

Consequently, if we define $p'\in\Delta(\Pi_{\mathsf{rns}})$ as the distribution induced by sampling $\pi\sim p$ and executing $\pi\circ_h\pi_{\mathrm{unif}}$, we have that $\Psi^M_{h,\varepsilon}(p')\le 5|\mathcal{A}|C^M_{\mathsf{push};h}\log(2\varepsilon^{-1})$.

Proof of Theorem 4.2. Let us abbreviate $\widetilde{d}_h^t=\sum_{i<t}d_h^{M,\pi^i}$. Observe that for $T=\frac{1}{\varepsilon}$, we have

$$\Psi^M_{\mathsf{push};h,\varepsilon}(p)=\sup_{\pi\in\Pi}\mathbb{E}^{M,\pi}\left[\frac{P_{h-1}^M(x_h\mid x_{h-1},a_{h-1})}{\mathbb{E}_{\pi'\sim p}\left[d_h^{M,\pi'}(x_h)\right]+\frac{1}{T}P_{h-1}^M(x_h\mid x_{h-1},a_{h-1})}\right]$$
$$= T\cdot\sup_{\pi\in\Pi}\mathbb{E}^{M,\pi}\left[\frac{P_{h-1}^M(x_h\mid x_{h-1},a_{h-1})}{\widetilde{d}_h^{T+1}(x_h)+P_{h-1}^M(x_h\mid x_{h-1},a_{h-1})}\right],$$

and hence it suffices to bound the quantity on the right-hand side. Observe that for all $t \in [T]$, we have that

$$\sup_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\widetilde{d}^t_h(x_h) + P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\right] \le \sup_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\widetilde{d}^{t-1}_h(x_h) + P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\right],$$

and consequently

$$\begin{aligned} T \cdot \sup_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\widetilde{d}^{T+1}_h(x_h) + P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\right] &\le \sum_{t=1}^T \sup_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\widetilde{d}^{t}_h(x_h) + P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\right] \\ &\le \sum_{t=1}^T \mathbb{E}^{M,\pi^t}\left[\frac{P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\widetilde{d}^{t}_h(x_h) + P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\right] + \varepsilon_{\mathrm{opt}} T. \end{aligned}$$

Now, let $\mu \in \Delta(\mathcal{X})$ attain the value of $C^M_{\mathsf{push};h}$. Using Lemma C.11 with $\varepsilon = 1$ and $\delta = C^M_{\mathsf{push};h}$, we have that for all $\pi \in \Pi$,

$$\begin{aligned} \mathbb{E}^{M,\pi}\left[\frac{P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\widetilde{d}^{t}_h(x_h) + P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\right] &\le \mathbb{E}^{M,\pi}\left[\frac{P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\widetilde{d}^{t}_h(x_h) + C^M_{\mathsf{push};h}\mu(x_h)}\right] + C^M_{\mathsf{push};h} \cdot \mathbb{E}^{M,\pi}\left[\frac{\mu(x_h)}{\widetilde{d}^{t}_h(x_h) + C^M_{\mathsf{push};h}\mu(x_h)}\right] \\ &\le 2C^M_{\mathsf{push};h} \cdot \mathbb{E}^{M,\pi}\left[\frac{\mu(x_h)}{\widetilde{d}^{t}_h(x_h) + C^M_{\mathsf{push};h}\mu(x_h)}\right]. \end{aligned}$$

Hence, we can bound

$$\begin{aligned} \sum_{t=1}^T \mathbb{E}^{M,\pi^t}\left[\frac{P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\widetilde{d}^{t}_h(x_h) + P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\right] &\le 2C^M_{\mathsf{push};h} \sum_{t=1}^T \mathbb{E}^{M,\pi^t}\left[\frac{\mu(x_h)}{\widetilde{d}^{t}_h(x_h) + C^M_{\mathsf{push};h}\mu(x_h)}\right] \\ &= 2C^M_{\mathsf{push};h} \sum_{x\in\mathcal{X}} \sum_{t=1}^T \mu(x) \frac{d^{M,\pi^t}(x)}{\widetilde{d}^{t}_h(x) + C^M_{\mathsf{push};h}\mu(x)}. \end{aligned}$$

Since $\sup_{\pi\in\Pi} d^{M,\pi}_h(x) \le \sup_{x'\in\mathcal{X}, a\in\mathcal{A}} P^M_{h-1}(x \mid x', a) \le C^M_{\mathsf{push};h}\mu(x)$ for all $x \in \mathcal{X}$, Lemma C.9 implies that

$$\sum_{x\in\mathcal{X},} \sum_{t=1}^T \mu(x) \frac{d^{M,\pi^t}(x)}{\widetilde{d}^{t}_h(x) + C_\infty \mu(x)} \le 2\log(2T).$$

We conclude that

$$\sum_{t=1}^T \mathbb{E}^{M,\pi^t}\left[\frac{P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\widetilde{d}^{t}_h(x_h) + P^M_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\right] \le 4C^M_{\mathsf{push};h}\log(2T) \tag{35}$$

and $\Psi^M_{\mathsf{push};h,\varepsilon}(p) \le 4C^M_{\mathsf{push};h}\log(2T) + \varepsilon_{\mathrm{opt}} T$. □

I. Proofs and Additional Details from Section 5

This section is organized as follows:

- Appendix I.1 presents our most general guarantee for Algorithm 2, Theorem I.1, and derives sample complexity bounds based on L_1-Coverability as a consequence.
- Appendix I.2 presents applications of these results to downstream policy optimization.
- Appendix I.3 presents preliminary technical lemmas.
- Appendix I.4 proves Theorem I.1, proving Theorem 5.1 as a corollary.

I.1. General Guarantees for Algorithm 2

In this section, we present general guarantees for CODEX (Algorithm 2) that (i) make us of online (as opposed to offline) estimation oracles, allowing for faster rates, and (ii) enjoy sample complexity scaling with L_1-Coverability, improving upon the L_∞-Coverability-based guarantees in Section 5.

I.1.1. Online Estimation Oracles

For *online estimation*, we measure the oracle's estimation performance in terms of cumulative Hellinger error, which we assume is bounded as follows.

Assumption I.1 (Online estimation oracle for $\mathcal{M}$)**.** *At each time $t \in [T]$, an online estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$ for $\mathcal{M}$ returns, given $\mathfrak{H}^{t-1} = (\pi^1, o^1), \ldots, (\pi^{t-1}, o^{t-1})$ with $o^i \sim M^\star(\pi^i)$ and $\pi^i \sim p^i$, an estimator $\widehat{M}^t \in \mathcal{M}$ such that whenever $M^\star \in \mathcal{M}$,*

$$\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(T) := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi^t), M^\star(\pi^t)\big)\right] \le \mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M}, T, \delta),$$

with probability at least $1-\delta$, where $\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M}, T, \delta)$ is a known upper bound.

See Section 4 of Foster et al. (2021) or Foster & Rakhlin (2023) for further background on online estimation. Algorithm 2 supports offline and online estimators, but is most straightforward to analyze for online estimators, and gives tighter sample complexity bounds in this case. The requirement in Assumption I.1 that the online estimator is *proper* (i.e., has $\widehat{M}^t \in \mathcal{M}$) is quite stringent, as generic online estimation algorithms (e.g., Vovk's aggregating algorithm) are improper, and proper algorithms are only known for specialized MDP classes such as tabular MDPs (see discussion in Foster et al. (2021)).[18] This contrasts with offline estimation, where most standard algorithms such as MLE are proper. As such, we present bounds based on online estimators as secondary results, with our bounds based on offline estimation serving as the main results.

Offline-to-online conversion. On the technical side, our interest in proper online estimation arises from the following structural result, which shows that whenever the L_1-Coverability parameter is bounded, any algorithm with low offline estimation error also enjoys low online estimation error (with polynomial loss in rate).

Lemma I.1 (Offline-to-online)**.** *Any offline estimator $\widehat{M}^t$ that satisfies Assumption 5.2 with estimation error bound $\mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(\mathcal{M}, T, \delta)$ satisfies Assumption I.1 with* $\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M}, T, \delta) \le \widetilde{O}\big(H\big(C^{M^\star}_{\mathsf{avg}}(1 + \mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(\mathcal{M}, T, \delta))\big)^{1/3} T^{2/3}\big)$.

Note that $C^{M^\star}_{\mathsf{avg}} \le \mathsf{Cov}^{M^\star}_0$; we leave an extension to $\mathsf{Cov}^{M^\star}_\varepsilon$ for $\varepsilon > 0$ to future work.
We also make use of a tighter offline-to-online lemma based on the (larger) L_∞-Coverability parameter $C^{M^\star}_\infty$.

Lemma I.2 (Xie et al. (2023))**.** *Any offline estimator $\widehat{M}^t$ that satisfies Assumption 5.2 with estimation error bound $\mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(\mathcal{M}, T, \delta)$ satisfies Assumption I.1 with* $\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M}, T, \delta) \le \widetilde{O}\big(H\big(C^{M^\star}_\infty T \cdot \mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(\mathcal{M}, T, \delta)\big)^{1/2} + HC^{M^\star}_\infty\big)$.

Both lemmas lead to a degradation in rate with respect to T, but lead to sublinear online estimation error whenever the offline estimation error bound is sublinear.

I.1.2. General Guarantees for Algorithm 2

Our most general guarantee for Algorithm 2, which assumes access to an online estimation oracle, is as follows.

Theorem I.1 (General guarantee for Algorithm 2)**.** *With parameters $T \in \mathbb{N}$, $C \ge 1$, and $\varepsilon > 0$ and an online estimation oracle satisfying Assumption I.1, whenever the optimization problem in Eq. (13) is feasible at every round, Algorithm 2 produces a policy covers $p_1, \ldots, p_H \in \Delta(\Pi)$ such that with probability at least $1-\delta$, $\forall h \in [H]$: $\Psi^{M^\star}_{h,\varepsilon}(p_h) \le$*

$$11HC + \frac{12}{\varepsilon}\sqrt{\frac{H^3 C \cdot \mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M}, T, \delta)}{T}} + \frac{8H}{\varepsilon^2}\frac{\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M}, T, \delta)}{T}. \tag{36}$$

Theorem 5.1 is derived by combining this result with Lemma I.2. The next result instantiates Theorem I.1 with Lemma I.1, allowing us to give sample complexity guarantees based on L_1-Coverability that support offline estimation oracles.

[18] On the statistical side, it is straightforward to extend the results in this section to accommodate improper online estimators; we impose this restriction for *computational* reasons, as this enables the application of the efficient planning results in Section 4.

Corollary I.1 (Main guarantee for Algorithm 2 under L_1-Coverability)**.** Let $\varepsilon > 0$ be given. Suppose that (i) we restrict $\mathcal{M}$ such that all $M \in \mathcal{M}$ have $\mathsf{Cov}_\varepsilon^M \leq \mathsf{Cov}_\varepsilon^{M^\star}$, and (ii) we solve Eq. (13) with $C = \mathsf{Cov}_\varepsilon^{M^\star}$ for all t (which is always feasible). Then, given access to an offline estimation oracle satisfying Assumptions 5.2 and 5.3, using $T = \widetilde{O}\Big(\frac{H^{12}(\mathsf{Cov}_0^{M^\star})^4 d_{\mathsf{est}} \log(B_{\mathsf{est}}/\delta)}{\varepsilon^6}\Big)$ episodes, Algorithm 2 produces policy covers $p_1, \ldots, p_H \in \Delta(\Pi)$ such that

$$\forall h \in [H]: \quad \Psi_{h,\varepsilon}^{M^\star}(p_h) \leq 12H \cdot \mathsf{Cov}_\varepsilon^{M^\star} \tag{37}$$

with probability at least $1-\delta$. In particular, for a finite class $\mathcal{M}$, if we use MLE as the estimator, we can take $T = \widetilde{O}\Big(\frac{H^{12}(\mathsf{Cov}_0^{M^\star})^4 \log(|\mathcal{M}|/\delta)}{\varepsilon^6}\Big)$.

This result shows that L_1-Coverability is itself a sufficiently powerful structural parameter to enable sample-efficient learning with nonlinear function approximation. Note that while Corollary I.1 assumes for simplicity that Eq. (13) is solved with $C = \mathsf{Cov}_\varepsilon^{M^\star}$, it should be clear that if we solve the objective for $C > \mathsf{Cov}_\varepsilon^{M^\star}$ the result continues to hold with $\mathsf{Cov}_\varepsilon^{M^\star}$ replaced by C in the sample complexity bound and approximation guarantee.

I.2. Applying Algorithm 2 to Downstream Policy Optimization

By Proposition 3.1 (see also Appendix D), the policy covers $p_1, \ldots, p_H$ returned by Algorithm 2 can be used to optimize any downstream reward function using standard offline RL algorithms. This leads to end-to-end guarantees for reward-driven PAC RL. For concreteness, we sketch an example which uses maximum likelihood (MLE) for offline policy optimization; see Appendix D for further examples and details.

Corollary I.2 (Application to reward-free reinforcement learning)**.** Given access to $H \cdot n$ trajectories from the policy covers $p_1, \ldots, p_H$ produced by Algorithm 2 (configured as in Theorem 5.1) and a realizable model class with $M^\star \in \mathcal{M}$, for any reward distribution $R = \{R_h\}_{h=1}^H$ with $\sum_{h=1}^H r_h \in [0,1]$, the Maximum Likelihood Estimation algorithm (described in Appendix D) produces a policy $\widehat{\pi}$ such that with probability at least $1-\delta$, $J_R^{M^\star}(\pi_\star) - J_R^{M^\star}(\widehat{\pi}) \leq O(H) \cdot \Big(\sqrt{HC_\infty^{M^\star} \cdot \frac{\log(|\mathcal{M}|/\delta)}{n}} + HC_\infty^{M^\star} \cdot \varepsilon\Big)$, where $J_R^{M^\star}(\pi) := \mathbb{E}^{M^\star,\pi}\Big[\sum_{h=1}^H r_h\Big]$ denotes the expected reward in $M^\star$ when R is the reward distribution.

We now sketch some basic examples in which Corollary I.2 can be applied.

Example I.1 (Tabular MDPs)**.** For tabular MDPs with $|\mathcal{X}| \leq S$ and $|\mathcal{A}| \leq A$, we can construct online estimators for which $\mathbf{Est}_{\mathsf{H}}^{\mathsf{on}}(\mathcal{M}, T, \delta) = \widetilde{O}(HS^2A)$, so that Theorem 5.1 gives sample complexity $T = \frac{\mathrm{poly}(H,S,A)}{\varepsilon^2}$ to compute policy covers such that $\Psi_{h,\varepsilon}^{M^\star}(p_h) \leq 12H \cdot C_\infty^{M^\star}$. ◁

Example I.2 (Low-Rank MDPs)**.** Consider the Low-Rank MDP model in Eq. (30) with dimension d and suppose, following Agarwal et al. (2020); Uehara et al. (2022), that we have access to classes Φ and Ψ such that $\phi_h \in \Phi$ and $\psi_h \in \Psi$. Then MLE achieves $\mathbf{Est}_{\mathsf{H}}^{\mathsf{off}}(\mathcal{M}, T, \delta) = \widetilde{O}(\log(|\Phi||\Psi|))$, and we can take $C_\infty^{M^\star} \leq d|\mathcal{A}|$, so Theorem 5.1 gives sample complexity $T = \frac{\mathrm{poly}(H,d,|\mathcal{A}|,\log(|\Phi||\Psi|))}{\varepsilon^4}$ to compute policy covers such that $\Psi_{h,\varepsilon}^{M^\star}(p_h) \leq 12H \cdot C_\infty^{M^\star}$. ◁

I.3. Technical Lemmas

Lemma I.1 (Offline-to-online)**.** *Any offline estimator $\widehat{M}^t$ that satisfies Assumption 5.2 with estimation error bound $\mathbf{Est}_{\mathsf{H}}^{\mathsf{off}}(\mathcal{M}, T, \delta)$ satisfies Assumption I.1 with* $\mathbf{Est}_{\mathsf{H}}^{\mathsf{on}}(\mathcal{M}, T, \delta) \leq \widetilde{O}\big(H\big(C_{\mathsf{avg}}^{M^\star}(1 + \mathbf{Est}_{\mathsf{H}}^{\mathsf{off}}(\mathcal{M}, T, \delta))\big)^{1/3} T^{2/3}\big)$.

Proof of Lemma I.1. Let us abbreviate $d_h^t = d_h^{M^\star,\pi^t}$. By Lemma A.11 of Foster et al. (2021), we have that

$$\begin{aligned}
\mathbf{Est}_{\mathsf{H}}^{\mathsf{on}}(T) &= \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t}\Big[D_{\mathsf{H}}^2\Big(M^\star(\pi^t), \widehat{M}^t(\pi^t)\Big)\Big] \\
&\leq O(\log(H)) \cdot \sum_{h=1}^H \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t} \mathbb{E}^{M^\star,\pi^t}\Big[D_{\mathsf{H}}^2\Big(P_h^{M^\star}(x_h,a_h), P_h^{\widehat{M}^t}(x_h,a_h)\Big) + D_{\mathsf{H}}^2\Big(R_h^{M^\star}(x_h,a_h), R_h^{\widehat{M}^t}(x_h,a_h)\Big)\Big].
\end{aligned}$$

At the same time, for all t, we have by Lemma C.5 that

$$\sum_{i<t} \mathbb{E}_{\pi^i \sim p^i} \mathbb{E}^{M^\star,\pi^i}\left[D^2_{\mathsf{H}}\left(P_h^{M^\star}(x_h,a_h), P_h^{\widehat{M}^t}(x_h,a_h)\right) + D^2_{\mathsf{H}}\left(R_h^{M^\star}(x_h,a_h), R_h^{\widehat{M}^t}(x_h,a_h)\right)\right]$$
$$\leq 4\sum_{i<t} \mathbb{E}_{\pi^i \sim p^i}\left[D^2_{\mathsf{H}}\left(M^\star(\pi^i), \widehat{M}^t(\pi^i)\right)\right] \leq 4\mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(t).$$

The result now follows by applying Lemma I.3—stated and proven below—to the function $g(x,a) = D^2_{\mathsf{H}}\big(P_h^{M^\star}(x,a), P_h^{\widehat{M}^t}(x,a)\big) + D^2_{\mathsf{H}}\big(R_h^{M^\star}(x,a), R_h^{\widehat{M}^t}(x,a)\big)$ for each layer $h \in [H]$, which has $B \leq 4$. □

Lemma I.2 (Xie et al. (2023))**.** *Any offline estimator $\widehat{M}^t$ that satisfies Assumption 5.2 with estimation error bound $\mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(\mathcal{M},T,\delta)$ satisfies Assumption I.1 with* $\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M},T,\delta) \leq \widetilde{O}\big(H\big(C^{M^\star}_\infty T \cdot \mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(\mathcal{M},T,\delta)\big)^{1/2} + HC^{M^\star}_\infty\big)$.

Proof of Lemma I.2. Let us abbreviate $d_h^t = d_h^{M^\star,\pi^t}$. By Lemma A.11 of Foster et al. (2021), we have that

$$\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(T) = \sum_{t=1}^T \mathbb{E}_{\pi^t\sim p^t}\left[D^2_{\mathsf{H}}\left(M^\star(\pi^t), \widehat{M}^t(\pi^t)\right)\right]$$
$$\leq O(\log(H)) \cdot \sum_{h=1}^H\sum_{t=1}^T \mathbb{E}_{\pi^t\sim p^t}\mathbb{E}^{M^\star,\pi^t}\left[D^2_{\mathsf{H}}\left(P_h^{M^\star}(x_h,a_h), P_h^{\widehat{M}^t}(x_h,a_h)\right) + D^2_{\mathsf{H}}\left(R_h^{M^\star}(x_h,a_h), R_h^{\widehat{M}^t}(x_h,a_h)\right)\right].$$

At the same time, for all t, we have by Lemma C.5 that

$$\sum_{i<t} \mathbb{E}_{\pi^i \sim p^i} \mathbb{E}^{M^\star,\pi^i}\left[D^2_{\mathsf{H}}\left(P_h^{M^\star}(x_h,a_h), P_h^{\widehat{M}^t}(x_h,a_h)\right) + D^2_{\mathsf{H}}\left(R_h^{M^\star}(x_h,a_h), R_h^{\widehat{M}^t}(x_h,a_h)\right)\right]$$
$$\leq 4\sum_{i<t} \mathbb{E}_{\pi^i \sim p^i}\left[D^2_{\mathsf{H}}\left(M^\star(\pi^i), \widehat{M}^t(\pi^i)\right)\right] \leq 4\mathbf{Est}^{\mathsf{off}}_{\mathsf{H}}(t).$$

The result now follows by applying Lemma I.4—stated and proven below—to the function $g(x,a) = D^2_{\mathsf{H}}\big(P_h^{M^\star}(x,a), P_h^{\widehat{M}^t}(x,a)\big) + D^2_{\mathsf{H}}\big(R_h^{M^\star}(x,a), R_h^{\widehat{M}^t}(x,a)\big)$ for each layer $h \in [H]$, which has $B \leq 4$. □

Lemma I.3. *Fix an MDP M and layer $h \in [H]$. Suppose we have a sequence of functions $g^1,\ldots,g^T \in [0,B]$ and policies $\pi^1,\ldots,\pi^T$ such that*

$$\forall t \in [T], \quad \sum_{i<t} \mathbb{E}^{M,\pi^i}\left[(g^t(x_h,a_h))^2\right] \leq \beta^2. \tag{38}$$

Then it holds that

$$\sum_{t=1}^T \mathbb{E}^{M,\pi^t}[g^t(x_h,a_h)] = 2(C^M_{\mathsf{avg};h})^{1/3}(\beta^2 B + B^3)^{1/3}T^{2/3}. \tag{39}$$

Proof of Lemma I.3. Let $\mu \in \Delta(\mathcal{X}\times\mathcal{A})$ denote a distribution that achieves the value $C^M_{\mathsf{avg};h}$. Let us abbreviate $d_h^t(x,a) = d_h^{M,\pi^t}(x,a)$ and

$$\widetilde{d}_h^t(x,a) = \sum_{i<t} d_h^{M,\pi^i}(x,a). \tag{40}$$

Observe that by Eq. (38) and the assumption that $g^t \in [0,B]$, we have that

$$\sum_{x\in\mathcal{X},a\in\mathcal{A}} \widetilde{d}_h^{t+1}(x,a)(g^t(x,a))^2 \leq \beta^2 + B^2 =: \alpha^2. \tag{41}$$

To begin, fix a parameter $\lambda > 0$ and define

$$\tau(x,a) := \min\Big\{t \mid \widetilde{d}_h^{t+1}(x,a) \geq \lambda\mu(x,a)\Big\}. \tag{42}$$

We can bound

$$\sum_{t=1}^{T} \mathbb{E}^{M,\pi^t}[g^t(x_h,a_h)] \leq \sum_{t=1}^{T} \mathbb{E}^{M,\pi^t}[g^t(x_h,a_h)\mathbb{I}\{t \geq \tau(x_h,a_h)\}] + B\sum_{t=1}^{T} \mathbb{E}^{M,\pi^t}[\mathbb{I}\{t < \tau(x_h,a_h)\}]. \tag{43}$$

For the second term above, we can write

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}^{M,\pi^t}[\mathbb{I}\{t < \tau(x_h,a_h)\}] &= \sum_{x,a}\sum_{t=1}^{T} d_h^t(x,a)\mathbb{I}\{t < \tau(x,a)\} \\ &= \sum_{x,a} \widetilde{d}_h^{\tau(x,a)}(x,a) < \lambda \sum_{x,a}\mu(x,a) = \lambda,\end{aligned}$$

where the final inequality uses the definition of $\tau(x,a)$.

For the first, term, using Cauchy-Schwarz, we can bound

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{E}^{M,\pi^t}[g^t(x_h,a_h)\mathbb{I}\{t \geq \tau(x_h,a_h)\}] \\ &= \sum_{t=1}^{T}\sum_{x\in\mathcal{X},a\in\mathcal{A}} d_h^t(x,a)g^t(x,a)\mathbb{I}\{t \geq \tau(x,a)\} \\ &= \sum_{t=1}^{T}\sum_{x\in\mathcal{X},a\in\mathcal{A}} \frac{d_h^t(x,a)}{(\widetilde{d}_h^{t+1}(x,a))^{1/2}}\mathbb{I}\{t \geq \tau(x,a)\}\cdot(\widetilde{d}_h^{t+1}(x,a))^{1/2}g^t(x,a) \\ &\leq \mathbf{A}^{1/2}\mathbf{B}^{1/2},\end{aligned}$$

where

$$\mathbf{A} := \sum_{t=1}^{T}\sum_{x\in\mathcal{X},a\in\mathcal{A}} \frac{(d_h^t(x,a))^2}{\widetilde{d}_h^{t+1}(x,a)}\mathbb{I}\{t \geq \tau(x,a)\}, \quad \text{and} \quad \mathbf{B} := \sum_{t=1}^{T}\sum_{x\in\mathcal{X},a\in\mathcal{A}} \widetilde{d}_h^{t+1}(x,a)(g^t(x,a))^2. \tag{44}$$

Eq. (41) implies that

$$\mathbf{B} = \sum_{t=1}^{T}\sum_{x\in\mathcal{X},a\in\mathcal{A}} \widetilde{d}_h^{t+1}(x,a)(g^t(x,a))^2 \leq \alpha^2 T.$$

It remains to bound term $\mathbf{A}$. From the definition of $\tau(x,a)$, we can bound

$$\mathbf{A} = \sum_{t=1}^{T}\sum_{x\in\mathcal{X},a\in\mathcal{A}} \frac{(d_h^t(x,a))^2}{\widetilde{d}_h^{t+1}(x,a)}\mathbb{I}\{t \geq \tau(x,a)\} \leq \frac{1}{\lambda}\sum_{t=1}^{T}\sum_{x\in\mathcal{X},a\in\mathcal{A}} \frac{(d_h^t(x,a))^2}{\mu(x,a)} \leq \frac{C^M_{\mathsf{avg};h}T}{\lambda},$$

where the last inequality uses that μ achieves the value of $C^M_{\mathsf{avg};h}$.

Combining the results so far, we have that

$$\sum_{t=1}^{T} \mathbb{E}^{M,\pi^t}[g^t(x_h,a_h)] \leq (\lambda^{-1}C^M_{\mathsf{avg};h}T)^{1/2}(\alpha^2 T)^{1/2} + \lambda B = (C^M_{\mathsf{avg};h}\alpha^2)^{1/2}T/\lambda^{1/2} + \lambda B.$$

We choose $\lambda = (C^M_{\mathsf{avg};h}\alpha^2)^{1/3}T^{2/3}/B^{2/3}$ to balance the terms, which gives a bound of the form

$$2(C^M_{\mathsf{avg};h}\alpha^2 B)^{1/3}T^{2/3} = 2(C^M_{\mathsf{avg};h})^{1/3}(\beta^2 B + B^3)^{1/3}T^{2/3}.$$

□

Lemma I.4. *Fix an MDP M and layer $h \in [H]$. Suppose we have a sequence of functions $g^1, \ldots, g^T \in [0, B]$ and policies $\pi^1, \ldots, \pi^T$ such that*

$$\forall t \in [T], \quad \sum_{i<t} \mathbb{E}^{M,\pi^i}\left[(g^t(x_h, a_h))^2\right] \leq \beta^2. \tag{45}$$

Then it holds that

$$\sum_{t=1}^{T} \mathbb{E}^{M,\pi^t}[g^t(x_h, a_h)] = O\Big(\sqrt{C^M_{\infty;h} T \log(T) \cdot \beta^2} + C^M_{\infty;h} B\Big). \tag{46}$$

Proof of Lemma I.4. See proof of Theorem 1 in Xie et al. (2023). □

I.4. Proofs from Section 5 and Appendix I.1

Theorem I.1 (General guarantee for Algorithm 2)**.** *With parameters $T \in \mathbb{N}$, $C \geq 1$, and $\varepsilon > 0$ and an online estimation oracle satisfying Assumption I.1, whenever the optimization problem in Eq. (13) is feasible at every round, Algorithm 2 produces a policy covers $p_1, \ldots, p_H \in \Delta(\Pi)$ such that with probability at least $1-\delta$, $\forall h \in [H]$: $\Psi^{M^\star}_{h,\varepsilon}(p_h) \leq$*

$$11HC + \frac{12}{\varepsilon}\sqrt{\frac{H^3 C \cdot \mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M}, T, \delta)}{T}} + \frac{8H}{\varepsilon^2}\frac{\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M}, T, \delta)}{T}. \tag{36}$$

Overview of proof. Before diving into the proof of Theorem I.1, we sketch the high-level idea. The crux of the analysis is to show that for each round t, we have that

$$\mathsf{Cov}^{M^\star}_{h,\varepsilon}(p^t_h) \lesssim \max_h \mathsf{Cov}^{\widehat{M}^t}_{h,\varepsilon}(p^t_h) + \frac{1}{\varepsilon}\sqrt{H^3 C \cdot \mathbb{E}_{\pi\sim q^t}\left[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi), M^\star(\pi)\big)\right]} + \frac{H}{\varepsilon^2} \cdot \mathbb{E}_{\pi\sim q^t}\left[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi), M^\star(\pi)\big)\right]. \tag{47}$$

Averaging across all rounds t, this allows us to conclude that

$$\frac{1}{T}\sum_{t=1}^{T} \mathsf{Cov}^{M^\star}_{h,\varepsilon}(p^t_h) \lesssim HC + \frac{1}{\varepsilon}\sqrt{H^3 C \cdot \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\pi\sim q^t}\left[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi), M^\star(\pi)\big)\right]} + \frac{H}{\varepsilon^2} \cdot \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\pi\sim q^t}\left[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi), M^\star(\pi)\big)\right].$$

Since $\mathsf{Cov}^{M^\star}_{h,\varepsilon}(p_h) \leq \frac{1}{T}\sum_{t=1}^{T} \mathsf{Cov}^{M^\star}_{h,\varepsilon}(p^t_h)$ and $\sum_{t=1}^{T} \mathbb{E}_{\pi\sim q^t}\left[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi), M^\star(\pi)\big)\right] \leq \mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M}, T, \delta)$ with probability at least $1-\delta$, this proves the result.

Eq. (47) can be thought as a reward-free analogue of a bound on the Decision-Estimation Coefficient (DEC) of Foster et al. (2021; 2023) and makes precise the reasoning that by optimizing the plug-in approximation to the L_1-Coverage objective, we either 1) cover the true MDP $M^\star$ well, or 2) achieve large information gain (as quantified by the instantaneous estimation error $\mathbb{E}_{\pi\sim q^t}\left[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi), M^\star(\pi)\big)\right]$).

Proof of Theorem I.1. Observe that by Jensen's inequality, we have that for all $h \in [H]$,

$$T \cdot \Psi^{M^\star}_{h,\varepsilon}(p_h) \leq \sum_{t=1}^{T} \Psi^{M^\star}_{h,\varepsilon}(p^t_h).$$

We will show how to bound the right-hand side above. To begin, we state two technical lemmas, both proven in the sequel.

Lemma I.5. *Fix $h \in [H]$. For all mixture policies $p \in \Delta(\Pi)$ and MDPs $\widehat{M} = \{\mathcal{X}, \mathcal{A}, \{P^{\widehat{M}}_h\}_{h=0}^H\}$, it holds that*

$$\Psi^{M^\star}_{h,\varepsilon}(p) \leq \max_\pi \mathbb{E}^{M^\star,\pi}\left[\frac{d^{M^\star,\pi}_h(x_h, a_h) + d^{\widehat{M},\pi}_h(x_h, a_h)}{d^{M^\star,p}_h(x_h, a_h) + \varepsilon \cdot (d^{M^\star,\pi}_h(x_h, a_h) + d^{\widehat{M},\pi}_h(x_h, a_h))}\right].$$

For the next result, for a given reward-free MDP $M = \{\mathcal{X}, \mathcal{A}, \{P^M_h\}_{h=0}^H\}$ and reward distribution $R = \{R_h\}_{h=1}^H$, we define

$$J^M_R(\pi) := \mathbb{E}^{M,\pi}\left[\sum_{h=1}^{H} r_h\right]$$

as the value under $r_h \sim R(x_h, a_h)$.

Lemma I.6. *Consider the reward-free setting. Let an MDP $\widehat{M} = \{\mathcal{X}, \mathcal{A}, \{P_h^{\widehat{M}}\}_{h=0}^H\}$ be given, and let $p_1, \ldots, p_H \in \Delta(\Pi)$ be (C, ε)-policy covers for $\widehat{M}$, i.e.*

$$\Psi_{h,\varepsilon}^{\widehat{M}}(p_h) \le C \quad \forall h \in [H]. \tag{48}$$

Then the distribution $q := \mathrm{Unif}(p_1, \ldots, p_H)$ ensures that for all MDPs $M = \{\mathcal{X}, \mathcal{A}, \{P_h^M\}_{h=0}^H\}$, all reward distributions $R = \{R_h\}_{h=1}^H$ with $\sum_{h=1}^H r_h \in [0, B]$ almost surely, and all policies $\pi \in \Pi$,

$$J_R^M(\pi) - J_R^{\widehat{M}}(\pi) \le 4B\sqrt{H^3 C \cdot \mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}\big(\widehat{M}(\pi), M(\pi)\big)\big]} + \sqrt{2}BHC \cdot \varepsilon.$$

For the remainder of the proof, we abbreviate $d_h^{M,\pi} \equiv d_h^{M,\pi}(x_h, a_h)$ whenever the argument is clear from context. Let $h \in [H]$ be fixed. We observe that by Lemma I.5,

$$\sum_{t=1}^T \Psi_{h,\varepsilon}^{M^\star}(p_h^t) \le \sum_{t=1}^T \max_{\pi\in\Pi} \mathbb{E}^{M^\star,\pi}\left[\frac{d_h^{M^\star,\pi}}{d_h^{M^\star,p_h^t} + \varepsilon\cdot(d_h^{M^\star,\pi} + d_h^{\widehat{M}^t,\pi})}\right] + \max_{\pi\in\Pi}\mathbb{E}^{M^\star,\pi}\left[\frac{d_h^{\widehat{M}^t,\pi}}{d_h^{M^\star,p_h^t} + \varepsilon\cdot(d_h^{M^\star,\pi} + d_h^{\widehat{M}^t,\pi})}\right].$$

For any $\pi \in \Pi$, by applying Lemma I.6 for each step t with the deterministic rewards

$$r_h^t(x,a) = \frac{d_h^{M^\star,\pi}(x,a)}{d_h^{M^\star,p^t}(x,a) + \varepsilon\cdot(d_h^{M^\star,\pi}(x,a) + d_h^{\widehat{M}^t,\pi}(x,a))}\mathbb{I}\{h' = h\},$$

which satisfy $\sum_{h=1}^H r_h^t \in \big[0, \varepsilon^{-1}\big]$ almost surely, we can bound

$$\begin{aligned}
&\mathbb{E}^{M^\star,\pi}\left[\frac{d_h^{M^\star,\pi}}{d_h^{M^\star,p_h^t} + \varepsilon\cdot(d_h^{M^\star,\pi} + d_h^{\widehat{M}^t,\pi})}\right] \\
&\le \mathbb{E}^{\widehat{M}^t,\pi}\left[\frac{d_h^{M^\star,\pi}}{d_h^{M^\star,p_h^t} + \varepsilon\cdot(d_h^{M^\star,\pi} + d_h^{\widehat{M}^t,\pi})}\right] + \frac{4}{\varepsilon}\sqrt{H^3C\cdot\mathbb{E}_{\pi\sim q^t}\big[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi), M^\star(\pi)\big)\big]} + \sqrt{2}HC, \\
&= \mathbb{E}^{M^\star,\pi}\left[\frac{d_h^{\widehat{M}^t,\pi}}{d_h^{M^\star,p_h^t} + \varepsilon\cdot(d_h^{M^\star,\pi} + d_h^{\widehat{M}^t,\pi})}\right] + \frac{4}{\varepsilon}\sqrt{H^3C\cdot\mathbb{E}_{\pi\sim q^t}\big[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi), M^\star(\pi)\big)\big]} + \sqrt{2}HC,
\end{aligned}$$

where we have simplified the second term by noting that $\sqrt{2}BHC\cdot\varepsilon = \sqrt{2}\varepsilon^{-1}HC\cdot\varepsilon = \sqrt{2}HC$. It follows that

$$\begin{aligned}
\sum_{t=1}^T \Psi_{h,\varepsilon}^{M^\star}(p_h^t) \le 2\sum_{t=1}^T \max_{\pi\in\Pi}\mathbb{E}^{M^\star,\pi}\left[\frac{d_h^{\widehat{M}^t,\pi}}{d_h^{M^\star,p_h^t} + \varepsilon\cdot(d_h^{M^\star,\pi} + d_h^{\widehat{M}^t,\pi})}\right] + \frac{4}{\varepsilon}\sum_{t=1}^T\sqrt{H^3C\cdot\mathbb{E}_{\pi\sim q^t}\big[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi), M^\star(\pi)\big)\big]} \\
+ \sqrt{2}HCT.
\end{aligned}$$

We now appeal to Lemma I.6 once more. For any π, by applying Lemma I.6 again at each step t with the rewards

$$r_h^t(x,a) = \frac{d_h^{\widehat{M}^t,\pi}(x,a)}{d_h^{M^\star,p^t}(x,a) + \varepsilon\cdot(d_h^{M^\star,\pi}(x,a) + d_h^{\widehat{M}^t,\pi}(x,a))}\mathbb{I}\{h' = h\},$$

which satisfy $\sum_{h=1}^H r_h \in [0, \varepsilon^{-1}]$ almost surely, allows us to bound

$$\begin{aligned}
&\mathbb{E}^{M^\star,\pi}\left[\frac{d_h^{\widehat{M}^t,\pi}}{d_h^{M^\star,p_h^t} + \varepsilon\cdot(d_h^{M^\star,\pi} + d_h^{\widehat{M}^t,\pi})}\right] \\
&\le \mathbb{E}^{\widehat{M}^t,\pi}\left[\frac{d_h^{\widehat{M}^t,\pi}}{d_h^{M^\star,p_h^t} + \varepsilon\cdot(d_h^{M^\star,\pi} + d_h^{\widehat{M}^t,\pi})}\right] + \frac{4}{\varepsilon}\sqrt{H^3C\cdot\mathbb{E}_{\pi\sim q^t}\big[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi), M^\star(\pi)\big)\big]} + \sqrt{2}HC, \\
&\le \mathbb{E}^{\widehat{M}^t,\pi}\left[\frac{d_h^{\widehat{M}^t,\pi}}{d_h^{M^\star,p_h^t} + \varepsilon\cdot d_h^{\widehat{M}^t,\pi}}\right] + \frac{4}{\varepsilon}\sqrt{H^3C\cdot\mathbb{E}_{\pi\sim q^t}\big[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi), M^\star(\pi)\big)\big]} + \sqrt{2}HC.
\end{aligned}$$

We conclude that

$$\sum_{t=1}^{T}\Psi^{M^\star}_{h,\varepsilon}(p^t_h) \le 2\sum_{t=1}^{T}\max_{\pi\in\Pi}\mathbb{E}^{\widehat{M}^t,\pi}\left[\frac{d^{\widehat{M}^t,\pi}_h}{d^{M^\star,p^t_h}_h+\varepsilon\cdot d^{\widehat{M}^t,\pi}_h}\right] + \frac{12}{\varepsilon}\sum_{t=1}^{T}\sqrt{H^3C\cdot\mathbb{E}_{\pi\sim q^t}\left[D^2_{\mathsf{H}}\big(\widehat{M}^t(\pi),M^\star(\pi)\big)\right]} + 3\sqrt{2}HCT.$$
$$\le 2\sum_{t=1}^{T}\max_{\pi\in\Pi}\mathbb{E}^{\widehat{M}^t,\pi}\left[\frac{d^{\widehat{M}^t,\pi}_h}{d^{M^\star,p^t_h}_h+\varepsilon\cdot d^{\widehat{M}^t,\pi}_h}\right] + \frac{12}{\varepsilon}\sqrt{H^3CT\cdot\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M},T,\delta)} + 3\sqrt{2}HCT. \tag{49}$$

We now appeal to the following lemma.

Lemma I.7. *Consider the reward-free setting. For any pair of MDP $\widehat{M}=\{\mathcal{X},\mathcal{A},\{P^{\widehat{M}}_h\}_{h=0}^H\}$ and $M=\{\mathcal{X},\mathcal{A},\{P^M_h\}_{h=0}^H\}$, any policy $\pi\in\Pi_{\mathsf{rns}}$, and any distribution $p\in\Delta(\Pi_{\mathsf{rns}})$, it holds that*

$$\mathbb{E}^{\widehat{M},\pi}\left[\frac{d^{\widehat{M},\pi}_h}{d^{M^\star,p}_h+\varepsilon\cdot d^{\widehat{M},\pi}_h}\right] \le 3\,\mathbb{E}^{\widehat{M},\pi}\left[\frac{d^{\widehat{M},\pi}_h}{d^{\widehat{M},p}_h+\varepsilon\cdot d^{\widehat{M},\pi}_h}\right] + \frac{4}{\varepsilon^2}\,\mathbb{E}_{\pi\sim p}\Big[D^2_{\mathsf{H}}\Big(\widehat{M}(\pi),M^\star(\pi)\Big)\Big].$$

Combining Eq. (49) with Lemma I.7, we have that

$$\sum_{t=1}^{T}\Psi^{M^\star}_{h,\varepsilon}(p^t_h) \le 6\sum_{t=1}^{T}\Psi^{\widehat{M}^t}_{h,\varepsilon}(p^t_h) + \frac{12}{\varepsilon}\sqrt{H^3CT\cdot\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M},T,\delta)} + 3\sqrt{2}HCT + \frac{8}{\varepsilon^2}\sum_{t=1}^{T}\mathbb{E}_{\pi\sim p^t_h}\Big[D^2_{\mathsf{H}}\Big(\widehat{M}^t(\pi),M^\star(\pi)\Big)\Big]$$
$$\le 6\sum_{t=1}^{T}\Psi^{\widehat{M}^t}_{h,\varepsilon}(p^t_h) + \frac{12}{\varepsilon}\sqrt{H^3CT\cdot\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M},T,\delta)} + 3\sqrt{2}HCT + \frac{8H}{\varepsilon^2}\sum_{t=1}^{T}\mathbb{E}_{\pi\sim q^t}\Big[D^2_{\mathsf{H}}\Big(\widehat{M}^t(\pi),M^\star(\pi)\Big)\Big]$$
$$\le 6\sum_{t=1}^{T}\Psi^{\widehat{M}^t}_{h,\varepsilon}(p^t_h) + \frac{12}{\varepsilon}\sqrt{H^3CT\cdot\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M},T,\delta)} + 3\sqrt{2}HCT + \frac{8H}{\varepsilon^2}\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M},T,\delta). \tag{50}$$

To conclude the proof of Eq. (36), we note that it follows from the definition of p^t_h that for all $h\in[H]$ and $t\in[T]$,

$$\Psi^{\widehat{M}^t}_{h,\varepsilon}(p^t_h)\le C.$$

Hence, we have that

$$\sum_{t=1}^{T}\Psi^{M^\star}_{h,\varepsilon}(p^t_h) \le 6CT + \frac{12}{\varepsilon}\sqrt{H^3CT\cdot\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M},T,\delta)} + 3\sqrt{2}HCT + \frac{8H}{\varepsilon^2}\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M},T,\delta)$$
$$\le 11HCT + \frac{12}{\varepsilon}\sqrt{H^3CT\cdot\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M},T,\delta)} + \frac{8H}{\varepsilon^2}\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M},T,\delta).$$

This implies that

$$\Psi^{M^\star}_{h,\varepsilon}(p_h) \le 11HC + \frac{12}{\varepsilon}\sqrt{\frac{H^3C\cdot\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M},T,\delta)}{T}} + \frac{8H}{\varepsilon^2}\frac{\mathbf{Est}^{\mathsf{on}}_{\mathsf{H}}(\mathcal{M},T,\delta)}{T}.$$

as desired.

□

Corollary I.1 (Main guarantee for Algorithm 2 under L_1-Coverability)**.** Let $\varepsilon>0$ be given. Suppose that (i) we restrict $\mathcal{M}$ such that all $M\in\mathcal{M}$ have $\mathsf{Cov}^M_\varepsilon\le\mathsf{Cov}^{M^\star}_\varepsilon$, and (ii) we solve Eq. (13) with $C=\mathsf{Cov}^{M^\star}_\varepsilon$ for all t (which is always feasible). Then, given access to an offline estimation oracle satisfying Assumptions 5.2 and 5.3, using $T=\widetilde{O}\Big(\frac{H^{12}(\mathsf{Cov}^{M^\star}_0)^4 d_{\mathsf{est}}\log(B_{\mathsf{est}}/\delta)}{\varepsilon^6}\Big)$ episodes, Algorithm 2 produces policy covers $p_1,\dots,p_H\in\Delta(\Pi)$ such that

$$\forall h\in[H]:\quad \Psi^{M^\star}_{h,\varepsilon}(p_h)\le 12H\cdot\mathsf{Cov}^{M^\star}_\varepsilon \tag{37}$$

with probability at least $1-\delta$. In particular, for a finite class $\mathcal{M}$, if we use MLE as the estimator, we can take $T=\widetilde{O}\Big(\frac{H^{12}(\mathsf{Cov}^{M^\star}_0)^4\log(|\mathcal{M}|/\delta)}{\varepsilon^6}\Big)$.

Proof of Corollary I.1. We prove Eq. (37) as a consequence of Theorem I.1. We can take $C \le \mathsf{Cov}_\varepsilon^{M^\star}$, and using Lemma I.1, we have

$$\mathbf{Est}_{\mathsf{H}}^{\mathsf{on}}(\mathcal{M},T,\delta) \le \widetilde{O}\Big(H\Big(C_{\mathsf{avg}}^{M^\star}(1 \vee \mathbf{Est}_{\mathsf{H}}^{\mathsf{off}}(\mathcal{M},T,\delta))\Big)^{1/3}T^{2/3}\Big) \le \widetilde{O}\Big(H\Big(\mathsf{Cov}_0^{M^\star}(1 \vee \mathbf{Est}_{\mathsf{H}}^{\mathsf{off}}(\mathcal{M},T,\delta))\Big)^{1/3}T^{2/3}\Big),$$

so that Eq. (36) gives

$$\forall h \in [H]: \quad \Psi_{h,\varepsilon}^{M^\star}(p_h) \le 11HC + \widetilde{O}\left(\frac{1}{\varepsilon}\left(\frac{H^{12}(\mathsf{Cov}_0^{M^\star})^4(1 \vee \mathbf{Est}_{\mathsf{H}}^{\mathsf{off}}(\mathcal{M},T,\delta))}{T}\right)^{1/6}\right) \le 12HC,$$

where the final inequality uses the choice for T in the corollary statement. □

Theorem 5.1 (Main guarantee for CODEX)**.** *Let $\varepsilon > 0$ be given. Let $C_\infty \equiv C_\infty^{M^\star}$, and suppose that (i) we restrict $\mathcal{M}$ such that all $M \in \mathcal{M}$ have $C_\infty^M \le C_\infty$, and (ii) we solve Eq. (13) with $C = C_\infty$ for all t.*[19] *Then, given an offline estimation oracle satisfying Assumptions 5.2 and 5.3, using $T = \widetilde{O}\Big(\frac{H^8(C_\infty^{M^\star})^3 d_{\mathsf{est}} \log(B_{\mathsf{est}}/\delta)}{\varepsilon^4}\Big)$ episodes, Algorithm 2 produces policy covers $p_1,\dots,p_H \in \Delta(\Pi)$ such that*

$$\forall h \in [H]: \quad \Psi_{h,\varepsilon}^{M^\star}(p_h) \le 12H \cdot C_\infty^{M^\star} \tag{15}$$

w.p. at least $1-\delta$. For a finite class $\mathcal{M}$, if we use MLE as the estimator, we can take $T = \widetilde{O}\Big(\frac{H^8(C_\infty^{M^\star})^3 \log(|\mathcal{M}|/\delta)}{\varepsilon^4}\Big)$.

Proof of Theorem 5.1. We prove Eq. (15) as a consequence of Theorem I.1. We can take $C \le C_\infty^{M^\star}$, and using Lemma I.2, we have

$$\mathbf{Est}_{\mathsf{H}}^{\mathsf{on}}(\mathcal{M},T,\delta) \le \widetilde{O}\Big(H\sqrt{C_\infty^{M^\star}T \cdot \mathbf{Est}_{\mathsf{H}}^{\mathsf{off}}(\mathcal{M},T,\delta)} + H \cdot C_\infty^{M^\star}\Big) \le \widetilde{O}\Big(H\sqrt{C_\infty^{M^\star}T \cdot (1 \vee \mathbf{Est}_{\mathsf{H}}^{\mathsf{off}}(\mathcal{M},T,\delta))}\Big),$$

so that Eq. (36) gives

$$\forall h \in [H]: \quad \Psi_{h,\varepsilon}^{M^\star}(p_h) \le 11HC + \widetilde{O}\left(\frac{1}{\varepsilon}\left(\frac{H^8(C_\infty^{M^\star})^3(1 \vee \mathbf{Est}_{\mathsf{H}}^{\mathsf{off}}(\mathcal{M},T,\delta))}{T}\right)^{1/4}\right) \le 12HC,$$

where the final inequality uses the choice for T in the theorem statement. □

I.4.1. Supporting Lemmas

Lemma I.5. *Fix $h \in [H]$. For all mixture policies $p \in \Delta(\Pi)$ and MDPs $\widehat{M} = \{\mathcal{X},\mathcal{A},\{P_h^{\widehat{M}}\}_{h=0}^H\}$, it holds that*

$$\Psi_{h,\varepsilon}^{M^\star}(p) \le \max_\pi \mathbb{E}^{M^\star,\pi}\left[\frac{d_h^{M^\star,\pi}(x_h,a_h) + d_h^{\widehat{M},\pi}(x_h,a_h)}{d_h^{M^\star,p}(x_h,a_h) + \varepsilon \cdot (d_h^{M^\star,\pi}(x_h,a_h) + d_h^{\widehat{M},\pi}(x_h,a_h))}\right].$$

Proof of Lemma I.5. To keep notation compact, let us suppress the dependence on x_h and a_h. For all $\pi \in \Pi_{\mathsf{rns}}$ and $p \in \Delta(\Pi_{\mathsf{rns}})$, we have that

$$\begin{aligned}
&\mathbb{E}^{M^\star,\pi}\left[\frac{d_h^{M^\star,\pi}}{d_h^{M^\star,p} + \varepsilon \cdot d_h^{M^\star,\pi}}\right] - \mathbb{E}^{M^\star,\pi}\left[\frac{d_h^{M^\star,\pi}}{d_h^{M^\star,p} + \varepsilon \cdot (d_h^{M^\star,\pi} + d_h^{\widehat{M},\pi})}\right] \\
&= \mathbb{E}^{M^\star,\pi}\left[\frac{d_h^{M^\star,\pi} \cdot \varepsilon \cdot d_h^{\widehat{M},\pi}}{(d_h^{M^\star,p} + \varepsilon \cdot d_h^{M^\star,\pi})(d_h^{M^\star,p} + \varepsilon \cdot (d_h^{M^\star,\pi} + d_h^{\widehat{M},\pi}))}\right] \\
&\le \mathbb{E}^{M^\star,\pi}\left[\frac{d_h^{\widehat{M},\pi}}{d_h^{M^\star,p} + \varepsilon \cdot (d_h^{M^\star,\pi} + d_h^{\widehat{M},\pi})}\right].
\end{aligned}$$

[19] We can take $C_\infty^M \le C_\infty$ w.l.o.g. when C_∞ is known. In this case, solving Eq. (13) with $C = C_\infty$ is feasible by Proposition 3.2.

This proves the result.

□

Lemma I.6. *Consider the reward-free setting. Let an MDP $\widehat{M} = \{\mathcal{X}, \mathcal{A}, \{P_h^{\widehat{M}}\}_{h=0}^H\}$ be given, and let $p_1, \ldots, p_H \in \Delta(\Pi)$ be (C, ε)-policy covers for $\widehat{M}$, i.e.*

$$\Psi_{h,\varepsilon}^{\widehat{M}}(p_h) \le C \quad \forall h \in [H]. \tag{48}$$

Then the distribution $q := \mathrm{Unif}(p_1, \ldots, p_H)$ ensures that for all MDPs $M = \{\mathcal{X}, \mathcal{A}, \{P_h^M\}_{h=0}^H\}$, all reward distributions $R = \{R_h\}_{h=1}^H$ with $\sum_{h=1}^H r_h \in [0, B]$ almost surely, and all policies $\pi \in \Pi$,

$$J_R^M(\pi) - J_R^{\widehat{M}}(\pi) \le 4B\sqrt{H^3 C \cdot \mathbb{E}_{\pi\sim q}\big[D_{\mathsf{H}}^2\big(\widehat{M}(\pi), M(\pi)\big)\big]} + \sqrt{2}BHC \cdot \varepsilon.$$

Proof of Lemma I.6. Let an arbitrary MDP $M = \{\mathcal{X}, \mathcal{A}, \{P_h^M\}_{h=0}^H\}$, reward distribution $R = \{R_h\}_{h=1}^H$, and policy $\pi \in \Pi$ be fixed. To begin, using the simulation lemma (Lemma C.8), we have

$$J_R^M(\pi) - J_R^{\widehat{M}}(\pi) \le B \cdot \sum_{h=1}^H \mathbb{E}^{\widehat{M},\pi}\Big[D_{\mathsf{H}}\big(P_h^{\widehat{M}}(x_h, a_h), P_h^M(x_h, a_h)\big)\Big],$$

where we have used that both MDPs have the same reward distribution. Let $h \in [H]$ be fixed. Since $D_{\mathsf{H}}\big(P_h^{\widehat{M}}(x_h, a_h), P_h^M(x_h, a_h)\big) \in [0, \sqrt{2}]$, we can use Proposition 3.1 to bound

$$\begin{aligned}
\mathbb{E}^{\widehat{M},\pi}\big[D_{\mathsf{H}}\big(P_h^{\widehat{M}}(x_h, a_h), P_h^M(x_h, a_h)\big)\big] &\le 2\sqrt{\Psi_{h,\varepsilon}^{\widehat{M}}(p_h) \cdot \mathbb{E}^{\widehat{M},p_h}\big[D_{\mathsf{H}}^2\big(P_h^{\widehat{M}}(x_h, a_h), P_h^M(x_h, a_h)\big)\big]} + \sqrt{2} \cdot \Psi_{h,\varepsilon}^{\widehat{M}}(p_h) \cdot \varepsilon \\
&\le 2\sqrt{H\Psi_{h,\varepsilon}^{\widehat{M}}(p_h) \cdot \mathbb{E}^{\widehat{M},q}\big[D_{\mathsf{H}}^2\big(P_h^{\widehat{M}}(x_h, a_h), P_h^M(x_h, a_h)\big)\big]} + \sqrt{2} \cdot \Psi_{h,\varepsilon}^{\widehat{M}}(p_h) \cdot \varepsilon \\
&\le 2\sqrt{HC \cdot \mathbb{E}^{\widehat{M},q}\big[D_{\mathsf{H}}^2\big(P_h^{\widehat{M}}(x_h, a_h), P_h^M(x_h, a_h)\big)\big]} + \sqrt{2}C \cdot \varepsilon \\
&\le 4\sqrt{HC \cdot \mathbb{E}_{\pi\sim q}\big[D_{\mathsf{H}}^2\big(\widehat{M}(\pi), M(\pi)\big)\big]} + \sqrt{2}C \cdot \varepsilon,
\end{aligned}$$

where the last inequality follows form Lemma C.5. Summing across all layers, we conclude that

$$J_R^M(\pi) - J_R^{\widehat{M}}(\pi) \le 4B\sqrt{H^3 C \cdot \mathbb{E}_{\pi\sim q}\big[D_{\mathsf{H}}^2\big(\widehat{M}(\pi), M(\pi)\big)\big]} + \sqrt{2}BHC \cdot \varepsilon.$$

□

Lemma I.7. *Consider the reward-free setting. For any pair of MDP $\widehat{M} = \{\mathcal{X}, \mathcal{A}, \{P_h^{\widehat{M}}\}_{h=0}^H\}$ and $M = \{\mathcal{X}, \mathcal{A}, \{P_h^M\}_{h=0}^H\}$, any policy $\pi \in \Pi_{\mathsf{rns}}$, and any distribution $p \in \Delta(\Pi_{\mathsf{rns}})$, it holds that*

$$\mathbb{E}^{\widehat{M},\pi}\left[\frac{d_h^{\widehat{M},\pi}}{d_h^{M^\star,p} + \varepsilon \cdot d_h^{\widehat{M},\pi}}\right] \le 3\,\mathbb{E}^{\widehat{M},\pi}\left[\frac{d_h^{\widehat{M},\pi}}{d_h^{\widehat{M},p} + \varepsilon \cdot d_h^{\widehat{M},\pi}}\right] + \frac{4}{\varepsilon^2}\,\mathbb{E}_{\pi\sim p}\Big[D_{\mathsf{H}}^2\Big(\widehat{M}(\pi), M^\star(\pi)\Big)\Big].$$

Proof of Lemma I.7. Consider any $c \geq 1$. For any $\pi \in \Pi_{\mathsf{rns}}$ and $p \in \Delta(\Pi_{\mathsf{rns}})$, we can write

$$\begin{aligned}
&\mathbb{E}^{\widehat{M},\pi}\left[\frac{d_h^{\widehat{M},\pi}}{d_h^{M^\star,p}+\varepsilon\cdot d_h^{\widehat{M},\pi}}\right] - c\cdot\mathbb{E}^{\widehat{M},\pi}\left[\frac{d_h^{\widehat{M},\pi}}{d_h^{\widehat{M},p}+\varepsilon\cdot d_h^{\widehat{M},\pi}}\right] \\
&= \mathbb{E}^{\widehat{M},\pi}\left[\frac{d_h^{\widehat{M},\pi}\left(d_h^{\widehat{M},p}+\varepsilon\cdot d_h^{\widehat{M},\pi} - c\cdot\left(d_h^{M^\star,p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)\right)}{\left(d_h^{M^\star,p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)\left(d_h^{\widehat{M},p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)}\right] \\
&\leq \mathbb{E}^{\widehat{M},\pi}\left[\frac{d_h^{\widehat{M},\pi}\left(d_h^{\widehat{M},p} - c\cdot d_h^{M^\star,p}\right)}{\left(d_h^{M^\star,p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)\left(d_h^{\widehat{M},p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)}\right] \\
&= \mathbb{E}^{\widehat{M},p}\left[\frac{(d_h^{\widehat{M},\pi})^2}{\left(d_h^{M^\star,p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)\left(d_h^{\widehat{M},p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)}\right] - c\cdot\mathbb{E}^{M^\star,p}\left[\frac{(d_h^{\widehat{M},\pi})^2}{\left(d_h^{M^\star,p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)\left(d_h^{\widehat{M},p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)}\right].
\end{aligned}$$

Observe that

$$\frac{(d_h^{\widehat{M},\pi})^2}{\left(d_h^{M^\star,p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)\left(d_h^{\widehat{M},p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)} \leq \frac{1}{\varepsilon^2}$$

almost surely. Consequently, Lemma C.6 implies that for $c = 3$,

$$\begin{aligned}
&\mathbb{E}^{\widehat{M},p}\left[\frac{(d_h^{\widehat{M},\pi})^2}{\left(d_h^{M^\star,p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)\left(d_h^{\widehat{M},p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)}\right] - 3\cdot\mathbb{E}^{M^\star,p}\left[\frac{(d_h^{\widehat{M},\pi})^2}{\left(d_h^{M^\star,p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)\left(d_h^{\widehat{M},p}+\varepsilon\cdot d_h^{\widehat{M},\pi}\right)}\right]. \\
&\leq \frac{4}{\varepsilon^2} D^2_{\mathsf{H}}\left(d_h^{\widehat{M},p}, d_h^{M^\star,p}\right).
\end{aligned}$$

Finally, note that by joint convexity of Hellinger distance and the data processing inequality, we have that

$$D^2_{\mathsf{H}}\left(d_h^{\widehat{M},p}, d_h^{M^\star,p}\right) \leq \mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(d_h^{\widehat{M},\pi}, d_h^{M^\star,\pi}\right)\right] \leq \mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(\widehat{M}(\pi), M^\star(\pi)\right)\right].$$

□

J. Proofs from Appendix F

Proposition F.1. *For all $\varepsilon > 0$, it holds that* $\mathsf{Cov}^M_{h,\varepsilon} \leq 1 + 2\sqrt{\frac{C^M_{\mathsf{avg};h}}{\varepsilon}}$.

Proof of Proposition F.1. Let $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$ be the distribution that attains the value of $C^M_{\mathsf{avg};h}$. Using Lemma C.10, we have that for all $\pi \in \Pi$ and $p \in \Delta(\Pi)$,

$$
\begin{aligned}
&\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon \cdot d_h^{M,\pi}(x_h,a_h)}\right] \\
&\leq 2\,\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon \cdot (d_h^{M,\pi}(x_h,a_h) + \mu(x_h,a_h))}\right] + \mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon \cdot (d_h^{M,\pi}(x_h,a_h) + \mu(x_h,a_h))}\right] \\
&\leq 2\,\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon \cdot (d_h^{M,\pi}(x_h,a_h) + \mu(x_h,a_h))}\right] + \mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon \cdot \mu(x_h,a_h))}\right].
\end{aligned}
$$

Observe that we can bound

$$
\begin{aligned}
&\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon \cdot (d_h^{M,\pi}(x_h,a_h) + \mu(x_h,a_h))}\right] \\
&= \sum_{x\in\mathcal{X},a\in\mathcal{A}} \frac{(d_h^{M,\pi}(x,a))^2}{d_h^{M,p}(x,a) + \varepsilon \cdot (d_h^{M,\pi}(x,a) + \mu(x,a))} \\
&= \sum_{x\in\mathcal{X},a\in\mathcal{A}} \frac{d_h^{M,\pi}(x,a)\mu^{1/2}(x,a)}{d_h^{M,p}(x,a) + \varepsilon \cdot (d_h^{M,\pi}(x,a) + \mu(x,a))} \cdot \frac{d_h^{M,\pi}(x,a)}{\mu^{1/2}(x,a)} \\
&\leq \left(\sum_{x\in\mathcal{X},a\in\mathcal{A}} \mu(x,a)\frac{(d_h^{M,\pi}(x,a))^2}{(d_h^{M,p}(x,a) + \varepsilon \cdot (d_h^{M,\pi}(x,a) + \mu(x,a)))^2}\right)^{1/2} \cdot \left(\sum_{x\in\mathcal{X},a\in\mathcal{A}} \frac{(d_h^{M,\pi}(x,a))^2}{\mu(x,a)}\right)^{1/2} \\
&\leq \left(\frac{1}{\varepsilon}\sum_{x\in\mathcal{X},a\in\mathcal{A}} \mu(x,a)\frac{d_h^{M,\pi}(x,a)}{d_h^{M,p}(x,a) + \varepsilon \cdot (d_h^{M,\pi}(x,a) + \mu(x,a))}\right)^{1/2} \cdot \left(C^M_{\mathsf{avg};h}\right)^{1/2} \\
&\leq \left(\frac{1}{\varepsilon}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon \cdot \mu(x_h,a_h)}\right]\right)^{1/2} \cdot \left(C^M_{\mathsf{avg};h}\right)^{1/2}.
\end{aligned}
$$

Hence, if we define

$$
V := \inf_{p\in\Delta(\Pi)} \sup_{\pi\in\Pi} \mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon \cdot \mu(x_h,a_h)}\right],
$$

this argument establishes that

$$
\mathsf{Cov}^M_{h,\varepsilon} \leq V + 2\sqrt{\frac{C^M_{\mathsf{avg};h}}{\varepsilon} \cdot V}.
$$

We now claim that $V \leq 1$. To see this, observe that the function

$$
(p,q) \mapsto \mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,p}(x_h,a_h) + \varepsilon \cdot \mu(x_h,a_h)}\right]
$$

is convex-concave. In addition, it is straightforward to see that the function is jointly Lipschitz with respect to total variation

distance whenever $\varepsilon > 0$. Hence, using the minimax theorem (Lemma C.1), we have that

$$\begin{aligned}
V &= \inf_{p\in\Delta(\Pi)} \sup_{q\in\Delta(\Pi)} \mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,p}(x_h,a_h)+\varepsilon\cdot\mu(x_h,a_h)}\right] \\
&= \sup_{q\in\Delta(\Pi)} \inf_{p\in\Delta(\Pi)} \mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,p}(x_h,a_h)+\varepsilon\cdot\mu(x_h,a_h)}\right] \\
&\le \sup_{q\in\Delta(\Pi)} \mathbb{E}_{\pi\sim q}\,\mathbb{E}^{M,\pi}\left[\frac{\mu(x_h,a_h)}{d_h^{M,q}(x_h,a_h)+\varepsilon\cdot\mu(x_h,a_h)}\right] \\
&= \sum_{x\in\mathcal{X},a\in\mathcal{A}} \frac{d_h^{M,q}(x,a)\mu(x,a)}{d_h^{M,q}(x,a)+\varepsilon\cdot\mu(x,a)} \le 1.
\end{aligned}$$

□

Proposition F.2. *Suppose the MDP M obeys the low-rank structure in Eq. (30). Then for all $h \in [H]$, we have $C^M_{\mathsf{avg};h} \le |\mathcal{A}|\cdot C^M_{\phi;h-1}$, and consequently $\mathsf{Cov}^M_{h,\varepsilon} \le 1 + 2\sqrt{\frac{|\mathcal{A}|\cdot C^M_{\phi;h-1}}{\varepsilon}}$.*

Proof of Proposition F.2. We first note that

$$C^M_{\mathsf{avg};h} \le |\mathcal{A}|\cdot \inf_{\mu\in\Delta(\mathcal{A})}\sup_{\pi\in\Pi}\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h)}{\mu(x_h)}\right].$$

Let $\nu \in \Delta(\mathcal{X}\times\mathcal{A})$ be arbitrary, and let $\mu := \nu \circ_{h-1} P^M_{h-1}$ be the distribution induced by sampling $(x_{h-1},a_{h-1}) \sim \nu$ and $x_h \sim P^M_{h-1}(\cdot \mid x_{h-1},a_{h-1})$. Let $\pi \in \Pi$ be arbitrary. We can write

$$\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h)}{\mu(x_h)}\right] = \left\langle \mathbb{E}^{M,\pi}[\phi_{h-1}(x_{h-1},a_{h-1})], \underbrace{\sum_{x\in\mathcal{X}}\psi_h(x)\frac{d_h^{M,\pi}(x)}{\mu(x)}}_{=:w}\right\rangle.$$

Using Cauchy-Schwarz and defining $\Sigma_\nu := \mathbb{E}_{(x_{h-1},a_{h-1})\sim\nu}\left[\phi_{h-1}(x_{h-1},a_{h-1})\phi_{h-1}(x_{h-1},a_{h-1})^\top\right]$, we can bound

$$\begin{aligned}
\langle \mathbb{E}^{M,\pi}[\phi_{h-1}(x_{h-1},a_{h-1})], w\rangle &= \left\langle \Sigma_\nu^{-1/2}\,\mathbb{E}^{M,\pi}[\phi_{h-1}(x_{h-1},a_{h-1})], \Sigma_\nu^{1/2} w\right\rangle \\
&\le \frac{1}{2}\|\mathbb{E}^{M,\pi}[\phi_{h-1}(x_{h-1},a_{h-1})]\|^2_{\Sigma_\nu^{-1}} + \frac{1}{2}\|w\|^2_{\Sigma_\nu}.
\end{aligned}$$

We can write

$$\begin{aligned}
\|w\|^2_{\Sigma_\nu} &= \mathbb{E}_{(x_{h-1},a_{h-1})\sim\nu}\left[\langle\phi_{h-1}(x_{h-1},a_{h-1}), w\rangle^2\right] \\
&= \mathbb{E}_{(x_{h-1},a_{h-1})\sim\nu}\left[\left\langle \phi_{h-1}(x_{h-1},a_{h-1}), \sum_{x\in\mathcal{X}}\psi_h(x)\frac{d_h^{M,\pi}(x)}{\mu(x)}\right\rangle^2\right] \\
&= \mathbb{E}_{(x_{h-1},a_{h-1})\sim\nu}\left[\left(\mathbb{E}^M\left[\frac{d_h^{M,\pi}(x_h)}{\mu(x_h)} \mid x_{h-1},a_{h-1}\right]\right)^2\right] \\
&\le \mathbb{E}_{x_h\sim\nu\circ P}\left[\left(\frac{d_h^{M,\pi}(x_h)}{\mu(x_h)}\right)^2\right] = \mathbb{E}_{x_h\sim\mu}\left[\left(\frac{d_h^{M,\pi}(x_h)}{\mu(x_h)}\right)^2\right] = \mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h)}{\mu(x_h)}\right].
\end{aligned}$$

Hence, we have shown that

$$\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h)}{\mu(x_h)}\right] \le \frac{1}{2}\|\mathbb{E}^{M,\pi}[\phi_{h-1}(x_{h-1},a_{h-1})]\|^2_{\Sigma_\nu^{-1}} + \frac{1}{2}\,\mathbb{E}^{M,\pi}\left[\frac{d_h^{M,\pi}(x_h)}{\mu(x_h)}\right].$$

To conclude, we rearrange and recall that 1) π is arbitrary, and 2) we are free to choose ν to minimize the right-hand side. From here, the claim follows from Proposition F.1. □

Proposition F.3. *There exists an MDP M and policy class $\Pi \subset \Pi_{\mathsf{rns}}$ with horizon $H = 1$ such that $C^M_{\infty;h} \leq 2$ (and hence $\mathsf{Cov}^M_{h,\varepsilon} \leq 2$ as well), yet for all $\varepsilon > 0$,*

$$\inf_{p \in \Delta(\Pi)} \Psi^M_{\infty;h,\varepsilon}(p) \geq \frac{1}{\varepsilon}, \tag{33}$$

and in particular $\inf_{p \in \Delta(\Pi)} \Psi^M_{\infty;h,0}(p) = \infty$.

Proof of Proposition F.3. Consider an MDP M with horizon $H = 1$, a singleton state space $\mathcal{S} = \{\mathfrak{s}\}$, and action space $\mathcal{A} = \mathbb{N} \cup \{\perp\}$. For each $i \in \mathbb{N}$, we define a randomized policy π^i via

$$\pi^i(a \mid \mathfrak{s}) = \begin{cases} 1 - \frac{1}{2i^2}, & a = \perp, \\ \frac{1}{2i^2}, & a = i, \\ 0, & \text{o.w.} \end{cases},$$

so that

$$d_1^{M,\pi^i}(\mathfrak{s}, a) = \begin{cases} 1 - \frac{1}{2i^2}, & z = \perp, \\ \frac{1}{2i^2}, & z = i, \\ 0, & \text{o.w.} \end{cases}$$

We set $\Pi = \{\pi^i\}_{i \in \mathbb{N}}$, and abbreviate $d^i(\mathfrak{s}, a) = d_1^{M,\pi^i}(\mathfrak{s}, a)$ going forward.

We first bound C^M_∞. We choose μ by setting $\mu(\mathfrak{s}, \perp) = \frac{1}{2}$ and $\mu(\mathfrak{s}, i) = \frac{3}{\pi^2} \cdot \frac{1}{i^2}$, which has $\sum_{a \in \mathcal{A}} \mu(\mathfrak{s}, a) = 1$. It is fairly immediate to see that for all i, we have $\frac{d^i(\mathfrak{s}, \perp)}{\mu(\mathfrak{s}, \perp)} \leq 2$ and

$$\frac{d^i(\mathfrak{s}, i)}{\mu(\mathfrak{s}, i)} = \frac{\pi^2}{6} \leq 2.$$

This shows that $C^M_\infty \leq 2$. On the other hand, for any $p \in \Delta(\mathbb{N})$, we have

$$\begin{aligned} \Psi^M_{\infty;h,\varepsilon}(p) \geq \sup_{i \in \mathbb{N}} \sup_{j \in \mathbb{N}} \left\{ \frac{d^i(\mathfrak{s}, j)}{\mathbb{E}_{k \sim p}[d^k(\mathfrak{s}, j)] + \varepsilon \cdot d^i(\mathfrak{s}, j)} \right\} &\geq \sup_{i \in \mathbb{N}} \frac{d^i(\mathfrak{s}, i)}{\mathbb{E}_{k \sim p}[d^k(\mathfrak{s}, i) + \varepsilon \cdot d^i(\mathfrak{s}, i)]} \\ &= \sup_{i \in \mathbb{N}} \frac{1/2i^2}{(p(i) + \varepsilon) \cdot (1/2i^2)} \\ &= \sup_{i \in \mathbb{N}} \frac{1}{p(i) + \varepsilon} = \frac{1}{\varepsilon}, \end{aligned}$$

where the conclusion holds because $\sum_{i \in \mathbb{N}} p(i) \leq 1$, which means for all $\delta > 0$, there exists i such that $p(i) \leq \delta$.

□

K. Proofs from Appendix E

K.1. General Guarantees for Algorithm 3

We first present general assumptions on the weight function estimation and policy optimization subroutines under which Algorithm 3 can be analyzed, then present our most general result, Theorem K.1.

K.1.1. Weight Function Realizability

Theorem E.1 is analyzed under the weight function realizability assumption in Assumption E.1. However, Algorithm 3 is most directly analyzed in terms of the following, slightly stronger weight function assumption, which we show is implied by Assumption E.1. To motivate the assumption, recall that we seek to estimate a weight function $\widehat{w}^t_h$ approximating Eq. (26).

Assumption K.1 (Weight function realizability—strong version)**.** *For a parameter $T \in \mathbb{N}$, we assume that for all $h \geq 2$, all $t \in [T]$, and all policies $\pi^1, \ldots, \pi^{t-1} \in \Pi_{\mathsf{ns}}$, we have that*

$$w_h^{\pi^1,\ldots,\pi^t}(x' \mid x, a) := \frac{P^{M^\star}_{h-1}(x' \mid x, a)}{\sum_{i<t} d_h^{M^\star,\pi^i}(x') + P^{M^\star}_{h-1}(x' \mid x, a)} \in \mathcal{W}_h.$$

We assume without loss of generality that $\|w\|_\infty \le 1$ *for all* $w \in \mathcal{W}_h$.

To compute a policy cover that approximately solves $p_h = \arg\min_{p \in \Delta(\Pi_{\mathsf{ns}})} \Psi^{M^\star}_{\mathsf{push};h,\varepsilon}(p_h)$ for parameter $\varepsilon > 0$, we require that Assumption K.1 holds for $T = \frac{1}{\varepsilon}$.

Algorithm 3 enjoys tighter guarantees when Assumption K.1 is satisfied, but the following result shows that Assumption E.1 implies Assumption K.1 at the cost of a small degradation in rate.

Proposition K.1. *For any* $T \in \mathbb{N}$*, given a weight function class* $\mathcal{W}$ *satisfying Assumption E.1, the induced class* $\mathcal{W}'$ *given by*

$$\mathcal{W}'_h := \left\{ (x,a,x') \mapsto \frac{1}{1 + \sum_{i<t} \frac{1}{w^i_h(x'|x,a)}} \mid w^1_h, \dots, w^{t-1}_h \in \mathcal{W}_h, t \in [T] \right\}$$

satisfies Assumption K.1, and has $\log|\mathcal{W}'_h| \le O(T \cdot \log|\mathcal{W}_h|)$.

For $T = \frac{1}{\varepsilon}$ this increases the weight function class size from $\log|\mathcal{W}|$ to $O\big(\frac{1}{\varepsilon} \cdot \log|\mathcal{W}|\big)$, leading to an extra $\frac{1}{\varepsilon}$ factor in the final sample complexity bound for our main result (Theorem E.1).

Remark K.1 (Sufficiency of Assumption K.1)**.** When invoked with layer $h \ge 2$ and iteration $t \ge 2$ within Algorithm 3, `EstimateWeightFunction` (Algorithm 4) collects datasets $\mathcal{D}_1 \sim \mu$ and $\mathcal{D}_2 \sim \nu$ such that

$$\mu(x' \mid x,a) = P^{M^\star}_{h-1}(x' \mid x,a), \quad \nu(x' \mid x,a) = \frac{1}{t}\left(\sum_{i<t} d^{M^\star,\pi^{h,i}}_h(x') + P^{M^\star}_{h-1}(x' \mid x,a) \right),$$

and

$$\mu(x,a) = \nu(x,a) = \frac{1}{2}\left(d^{M^\star,p_{h-1}}_{h-1}(x,a) + \frac{1}{t-1} \sum_{i<t} d^{M^\star,\pi^{h,i} \circ_{h-1} \pi_{\mathrm{unif}}}_{h-1}(x,a) \right). \tag{51}$$

Then, in Line 10, the algorithm computes the estimator Eq. (27) with respect to the class $t \cdot \mathcal{W}$, which is guaranteed to have $\frac{\mu(x,a,x')}{\nu(x,a,x')} = t \cdot w^t_h(x' \mid x,a) \in t \cdot \mathcal{W}$ under Assumption K.1.

K.1.2. Policy Optimization Subroutine

This section presents general conditions for the subroutine `PolicyOptimization` under which `CODEX.W` obtains the same guarantees as in Theorem E.1, and establishes that `PSDP` satisfies this assumption.

To formalize the requirement of `PolicyOptimization`, recall that for each layer $h \ge 2$, iteration $t \in [T]$, and each $\ell \le h-1$, we define

$$Q^{M^\star,\pi}_\ell(x,a;\widehat{w}^t_h) = \mathbb{E}^{M^\star,\pi}[\widehat{w}^t_h(x_h \mid x_{h-1},a_{h-1}) \mid x_\ell = x, a_\ell = a]$$

as the Q-function for a policy $\pi \in \Pi_{\mathsf{ns}}$ under the (stochastic) reward $r^t_{h-1} = \widehat{w}^t_h(x_h \mid x_{h-1},a_{h-1})$ in Algorithm 3. We assume that the policy $\pi^{t,h} = \mathtt{PolicyOptimization}_{h-1}(r^{h,t}, p_{1:h-1}, \epsilon, \delta)$ approximately maximizes this Q-function under $p_1, \dots, p_{h-1}$.

Assumption K.2 (Local optimality for policy optimization)**.** *For any fixed iteration* $h \ge 2$ *and* $t \in [T]$*, the subroutine* $\mathtt{PolicyOptimization}_{h-1}(r^{h,t}, p_{1:h-1}, \epsilon, \delta)$ *produces a policy* $\pi^{h,t}$ *such that with probability at least* $1-\delta$*,*

$$\sum_{\ell=1}^{h-1} \mathbb{E}^{M^\star,p_\ell}\left[\max_{a \in \mathcal{A}} Q^{M^\star,\pi^{h,t}}_\ell(x_\ell,a;\widehat{w}^t_h) - Q^{M^\star,\pi^{h,t}}_\ell(x_\ell,\pi^{h,t}(x_\ell);\widehat{w}^t_h) \right] \le \epsilon, \tag{52}$$

and does so using $N_{\mathsf{opt}}(\epsilon,\delta)$ *episodes.*

This assumption asserts that $\pi^{t,h}$ cannot be substantially improved, but only with respect to the state distribution induced by $p_1, \dots, p_{h-1}$.[20] This is a weak guarantee that can be achieved using only data collected from $p_1, \dots, p_{h-1}$ (e.g., via offline RL methods or hybrid offline/online methods), and does not require systematic exploration.

[20] Note that if $p_1, \dots, p_{h-1}$ uniformly cover all policies, then Assumption K.2 implies that $\pi^{h,t}$ is globally optimal by the performance difference lemma. However, Eq. (52) can still lead to useful guarantees in the presence of partial coverage, which our analysis critically exploits.

The next result shows that PSDP satisfies Assumption K.2 under the value function realizability assumption in Assumption E.2.

Lemma K.1 (Local optimality for PSDP)**.** *Suppose Assumption E.2 holds. Then for any $\epsilon, \delta \in (0,1)$, the subroutine $\pi^{h,t} = \mathsf{PSDP}_{h-1}(r^{h,t}, p_{1:h-1}, \epsilon, \delta)$ satisfies Assumption K.2, and does so using at most $N_{\mathsf{psdp}}(\epsilon, \delta) = O\Big(\frac{H^3|\mathcal{A}|\log(|\mathcal{Q}|H\delta^{-1})}{\epsilon^2}\Big)$ episodes.*

See Appendix K.4 for details. We expect that similar guarantees can be proven for Natural Policy Gradient, Conservative Policy Iteration, and other standard local search methods. Different subroutines may allow one to make use of weaker function approximation requirements.

K.1.3. General Guarantee for CODEX.W (Algorithm 3)

Our most general guarantee for CODEX.W is given below.

Theorem K.1 (General guarantee for Algorithm 3)**.** *Let $\varepsilon \in (0, 1/2)$ and $\delta \in (0, e^{-1})$ be given, and suppose that Assumption K.1 and Assumption K.2 are satisfied. Then Algorithm 3 produces policy covers $p_1, \ldots, p_H \in \Delta(\Pi_{\mathsf{rns}})$ such that with probability at least $1-\delta$, for all $h \in [H]$,*

$$\Psi^{M^\star}_{\mathsf{push};h,\varepsilon}(p_h) \leq 170H\log(\varepsilon^{-1}) \cdot C^{M^\star}_{\mathsf{push}},$$

and does so using at most

$$N \leq \widetilde{O}\Big(\frac{H|\mathcal{A}|\log(|\mathcal{W}|\delta^{-1})}{\varepsilon^3} + \frac{H}{\varepsilon}N_{\mathsf{opt}}(c\varepsilon^2, \delta/2HT)\Big)$$

episodes, where $c > 0$ is a sufficiently small absolute constant.

In particular, this result shows that we can optimize the pushforward coverability objective (and consequently the L_1-Coverage objective, via Proposition 4.2), up to small $O(H\log(\varepsilon^{-1}))$ approximation factor. The sample complexity is polynomial in all relevant problem parameters whenever the subroutine PolicyOptimization has polynomial sample complexity. Note that the sample complexity for the first term is of order $1/\varepsilon^3$ (as opposed to the slower $1/\varepsilon^4$ in Theorem E.1) since we are stating this result under the stronger weight function realizability assumption (Assumption K.1).

Combining Theorem K.1 with Proposition K.1 and the guarantee for PSDP (Lemma K.1), we obtain Theorem E.1.

K.2. Technical Preliminaries

Lemma K.2. *For any distribution $\omega \in \Delta(\mathcal{Z})$ and any pair of functions $w, w' : \mathcal{Z} \to \mathbb{R}_+$,*

$$\mathbb{E}_\omega[w] \leq 3\,\mathbb{E}_\omega[w'] + 2\,\mathbb{E}_\omega\big[(\sqrt{w} - \sqrt{w'})^2\big].$$

Proof of Lemma K.2. By AM-GM, we have

$$\begin{aligned}
|\mathbb{E}_\omega[w] - \mathbb{E}_\omega[w']| &\leq \mathbb{E}_\omega\Big[|\sqrt{w} - \sqrt{w'}|(\sqrt{w} + \sqrt{w'})\Big] \\
&\leq \sqrt{\mathbb{E}_\omega\big[(\sqrt{w} + \sqrt{w'})^2\big] \cdot \mathbb{E}_\omega\big[(\sqrt{w} - \sqrt{w'})^2\big]} \\
&\leq \frac{1}{2}(\mathbb{E}_\omega[w] + \mathbb{E}_\omega[w']) + \frac{1}{2}\,\mathbb{E}_\omega\big[(\sqrt{w} - \sqrt{w'})^2\big].
\end{aligned}$$

Rearranging, we conclude that

$$\mathbb{E}_\omega[w] \leq 3\,\mathbb{E}_\omega[w'] + 2\,\mathbb{E}_\omega\big[(\sqrt{w} - \sqrt{w'})^2\big].$$

□

Lemma K.3 (e.g., Xie et al. (2023); Mhammedi et al. (2023a))**.** *Consider a set $\mathcal{Z}$ and a sequence of distributions $d^1, \dots, d^T \in \Delta(\mathcal{Z})$ for which there exists a distribution $\mu \in \Delta(\mathcal{Z})$ such that $\sup_{z\in\mathcal{Z}}\left\{\frac{d^t(z)}{\mu(z)}\right\} \le C$ for all $t \in [T]$. For any sequence of functions $g^1, \dots, g^T \subset (\mathcal{Z} \to [-B, B])$, it holds that*

$$\sum_{t=1}^{T} \mathbb{E}_{z\sim d^t}[g(z)] \le \sqrt{2C\log(2T)\sum_{t=1}^{T}\sum_{i<t}\mathbb{E}_{z\sim d^i}[(g^t(z))^2]} + 2CB. \tag{53}$$

K.3. Weight Function Estimation

In this section, we give self-contained guarantees for the statistical problem of estimating the density ratio (or, "weight function") for a pair of distributions.

Consider the following setting. Let $\mathcal{Z}$ be a set. We receive samples $z_\mu^1, \dots, z_\mu^n \in \mathcal{Z}$ and $z_\nu^1, \dots, z_\nu^n \in \mathcal{Z}$, where $z_\mu^t \sim \mu^t \in \Delta(\mathcal{Z})$ and $z_\nu^t \sim \nu^t \in \Delta(\mathcal{Z})$. The distributions μ^t and ν^t can be chosen in an adaptive fashion based on $z_\mu^1, z_\nu^1, \dots, z_\mu^{t-1}, z_\nu^{t-1}$. We define $\mu = \frac{1}{n}\sum_{t=1}^n \mu^t$ and $\nu = \frac{1}{n}\sum_{t=1}^n \nu^t$, and our goal is to estimate the density ratio

$$w^\star(z) := \frac{\mu(z)}{\nu(z)}.$$

We assume that $\|w_\star\|_\infty \le B$, and assume access to a *realizable* weight function class $\mathcal{W}$ with $w^\star \in \mathcal{W}$. Following Nguyen et al. (2010) (see also Katdare et al. (2023)), we consider the estimator

$$\widehat{w} := \arg\max_{w\in\mathcal{W}} \widehat{\mathbb{E}}_\mu[\log(w)] - \widehat{\mathbb{E}}_\nu[w], \tag{54}$$

where $\widehat{\mathbb{E}}_\mu[\cdot]$ denotes the empirical expectation with respect to $z_\mu^1, \dots, z_\mu^n$ and $\widehat{\mathbb{E}}_\nu[\cdot]$ denotes the empirical expectation with respect to $z_\nu^1, \dots, z_\nu^n$. The following theorem gives a finite-sample bound for this estimator, which may be of independent interest.

Theorem K.2. *Suppose that $w^\star \in \mathcal{W}$ and $\sup_{w\in\mathcal{W}}\|w\|_\infty \le B$. The estimator in Eq. (54) ensures that with probability at least $1-\delta$,*

$$D^2_{\mathsf{H},\nu}(\widehat{w}, w^\star) \le \frac{20B\log(|\mathcal{W}|\delta^{-1})}{n},$$

where $D^2_{\mathsf{H},\nu}(w, w') := \mathbb{E}_\nu\left[\left(\sqrt{w} - \sqrt{w'}\right)^2\right]$.

Remark K.2 (Extension to contextual weight function estimation)**.** An immediate corollary for Theorem K.2 concerns the following "contextual" setting. Suppose that $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and that for all t, $z_\mu^t = (x^t, y_\mu^t)$ and $z_\nu = (x^t, y_\nu^t)$ have the same marginal distribution for x^t, i.e. $\mu^t(x, y) = \mu^t(y \mid x)\omega(x)$ and $\nu^t(x, y) = \nu^t(y \mid x)\omega(x)$ for some $\omega \in \Delta(\mathcal{X})$. Define $\mu(y \mid x) = \frac{1}{n}\sum_{t=1}^n \mu^t(y \mid x)$ and $\nu(y \mid x) = \frac{1}{n}\sum_{t=1}^n \nu^t(y \mid x)$, and let $w^\star(y \mid x) = \frac{\mu(y|x)}{\nu(y|x)}$. Then, given a class of weight functions $\mathcal{W}$ with $w^\star \in \mathcal{W}$, where each $w \in \mathcal{W}$ has the form $w(y \mid x)$ and $\|w\|_\infty \le B$, the estimator in Eq. (54) ensures that

$$\mathbb{E}_{x\sim\omega, y\sim\nu(\cdot|x)}[|\widehat{w}(y \mid x) - w^\star(y \mid x)|] \le 10B\sqrt{\frac{\log(|\mathcal{W}|\delta^{-1})}{n}}.$$

Proof of Theorem K.2. Define $V(w) = \mathbb{E}_\mu[\log(w)] - \mathbb{E}_\nu[w]$ and $\widehat{V}(w) = \widehat{\mathbb{E}}_\mu[\log(w)] - \widehat{\mathbb{E}}_\nu[w]$, and note that $w^\star = \frac{\mu}{\nu} = \arg\max_w V(w)$. We begin by performing concentration on the log-loss terms. Define $X_t(w) = \frac{1}{2}\big(\log(w^\star(z_\mu^t)) - \log(w(z_\mu^t))\big)$. By Lemma C.2 and a union bound, we have that with probability at least $1-\delta$, for all $w \in \mathcal{W}$,

$$\frac{1}{n}\sum_{t=1}^n -\log\left(\mathbb{E}_{\mu^t}\left[\exp\left(\frac{1}{2}\log(w/w^\star)\right)\right]\right) \le \frac{1}{2}\left(\widehat{\mathbb{E}}_\mu[\log(w^\star)] - \widehat{\mathbb{E}}_\mu[\log(w)]\right) + \frac{\log(|\mathcal{W}|\delta^{-1})}{n}.$$

Note that $\mathbb{E}_{\mu^t}\left[\exp\left(\frac{1}{2}\log(w/w^\star)\right)\right] = \mathbb{E}_{\mu^t}\left[\sqrt{w/w^\star}\right]$. We now state and prove a basic technical lemma.

Lemma K.4. *For all $x > 0$, $-\log(x) \geq 1 - x$.*

Proof of Lemma K.4. Let $f(x) = -\log(x)$. Since f is convex, we have that for all $x > 0$,

$$f(x) \geq f(1) + f'(1)(x-1).$$

Noting that $f(1) = 0$ and $f'(1) = -1$, the result is established. □

Using Lemma K.4, we have

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n} -\log\left(\mathbb{E}_{\mu^t}\left[\exp\left(\frac{1}{2}\log(w/w^\star)\right)\right]\right) &\geq 1 - \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\mu^t}\left[\sqrt{w/w^\star}\right] \\ &= 1 - \mathbb{E}_{\mu}\left[\sqrt{w/w^\star}\right] = 1 - \mathbb{E}_{\nu}\left[\sqrt{w \cdot w^\star}\right],\end{aligned}$$

where the last line uses that $\mathbb{E}_{\mu}\left[\sqrt{w/w^\star}\right] = \mathbb{E}_{\nu}\left[w^\star\sqrt{w/w^\star}\right] = \mathbb{E}_{\nu}\left[\sqrt{w \cdot w^\star}\right]$. By direct calculation, we have that

$$D^2_{\mathsf{H},\nu}(w, w^\star) = \mathbb{E}_{\nu}\left[\left(\sqrt{w} - \sqrt{w^\star}\right)^2\right] = \mathbb{E}_{\nu}[w^\star] + \mathbb{E}_{\nu}[w] - 2\,\mathbb{E}_{\nu}\left[\sqrt{w \cdot w^\star}\right] = 1 + \mathbb{E}_{\nu}[w] - 2\,\mathbb{E}_{\nu}\left[\sqrt{w \cdot w^\star}\right],$$

so that

$$1 - \mathbb{E}_{\nu}\left[\sqrt{w \cdot w^\star}\right] = \frac{1}{2}D^2_{\mathsf{H},\nu}(w, w^\star) + \frac{1}{2}(1 - \mathbb{E}_{\nu}[w]).$$

Specializing to $\widehat{w}$, we have

$$\begin{aligned}\frac{1}{2}D^2_{\mathsf{H},\nu}(\widehat{w}, w^\star) + \frac{1}{2}(1 - \mathbb{E}_{\nu}[\widehat{w}]) &\leq \frac{1}{2}\left(\widehat{\mathbb{E}}_{\mu}[\log(w^\star)] - \widehat{\mathbb{E}}_{\mu}[\log(\widehat{w})]\right) + \log(|\mathcal{W}|\delta^{-1}) \\ &= \frac{1}{2}\left(\widehat{V}(w^\star) - \widehat{V}(\widehat{w})\right) + \frac{1}{2}\left(\widehat{\mathbb{E}}_{\nu}[w^\star] - \widehat{\mathbb{E}}_{\nu}[\widehat{w}]\right) + \frac{\log(|\mathcal{W}|\delta^{-1})}{n}.\end{aligned}$$

Since $\widehat{V}(w^\star) - \widehat{V}(\widehat{w}) \leq 0$, rearranging gives

$$D^2_{\mathsf{H},\nu}(\widehat{w}, w^\star) \leq \left(\widehat{\mathbb{E}}_{\nu}[w^\star] - \widehat{\mathbb{E}}_{\nu}[\widehat{w}]\right) - \left(1 - \widehat{\mathbb{E}}_{\nu}[\widehat{w}]\right) + \frac{2\log(|\mathcal{W}|\delta^{-1})}{n} \tag{55}$$

$$= \left(\widehat{\mathbb{E}}_{\nu}[w^\star] - \widehat{\mathbb{E}}_{\nu}[\widehat{w}]\right) - (\mathbb{E}_{\nu}[w^\star] - \mathbb{E}_{\nu}[\widehat{w}]) + \frac{2\log(|\mathcal{W}|\delta^{-1})}{n}. \tag{56}$$

Using Lemma C.3, we have that for all $\eta \leq 1/2B$, with probability at least $1 - \delta$, for all $w \in \mathcal{W}$

$$\begin{aligned}\left(\widehat{\mathbb{E}}_{\nu}[w^\star] - \widehat{\mathbb{E}}_{\nu}[w]\right) - (\mathbb{E}_{\nu}[w^\star] - \mathbb{E}_{\nu}[w]) &\leq \frac{\eta}{n}\sum_{t=1}^{n}\mathbb{E}_{\nu_t}\left[(w - w^\star)^2\right] + \frac{\log(|\mathcal{W}|\delta^{-1})}{\eta n} \\ &= \eta\,\mathbb{E}_{\nu}\left[(w - w^\star)^2\right] + \frac{\log(|\mathcal{W}|\delta^{-1})}{\eta n}.\end{aligned}$$

We further observe that

$$\mathbb{E}_{\nu}\left[(w - w^\star)^2\right] = \mathbb{E}_{\nu}\left[(\sqrt{w} - \sqrt{w^\star})^2(\sqrt{w} + \sqrt{w^\star})^2\right] \leq 4B\,\mathbb{E}_{\nu}\left[(\sqrt{w} - \sqrt{w^\star})^2\right] = 4B \cdot D^2_{\mathsf{H},\nu}(w, w^\star),$$

so choosing $\eta = \frac{1}{8B}$ gives

$$\left(\widehat{\mathbb{E}}_{\nu}[w^\star] - \widehat{\mathbb{E}}_{\nu}[w]\right) - (\mathbb{E}_{\nu}[w^\star] - \mathbb{E}_{\nu}[w]) \leq \frac{1}{2}D^2_{\mathsf{H},\nu}(w, w^\star) + \frac{8B\log(|\mathcal{W}|\delta^{-1})}{n}. \tag{57}$$

Combining this with Eq. (55), we conclude that

$$D^2_{\mathsf{H},\nu}(\widehat{w}, w^\star) \leq \frac{1}{2}D^2_{\mathsf{H},\nu}(\widehat{w}, w^\star) + \frac{8B\log(|\mathcal{W}|\delta^{-1})}{n} + 2\frac{\log(|\mathcal{W}|\delta^{-1})}{n},$$

which implies that

$$D^2_{\mathsf{H},\nu}(\widehat{w}, w^\star) \le \frac{20B\log(|\mathcal{W}|\delta^{-1})}{n},$$

after simplifying. □

K.4. Policy Optimization Subroutines

K.4.1. Policy Search by Dynamic Programming (PSDP)

Algorithm 6 $\mathsf{PSDP}_h(p_{1:h}, r_{1:h}; \epsilon, \delta, \mathcal{Q}_{1:h})$: Policy Search by Dynamic Programming (cf. Bagnell et al. (2003))

1: **input:**
- Target layer $h \in [H]$, policy covers $p_{1:h}$, reward functions $r_{1:h}$.
- Accuracy parameters $\epsilon, \delta \in [0,1]$.
- Function classes $\mathcal{Q}_{1:h}$.

2: Let $n = n_{\mathsf{psdp}}(\epsilon, \delta) := c \cdot \frac{H^2|\mathcal{A}|\log(|\mathcal{Q}|H\delta^{-1})}{\epsilon^2}$ for a sufficiently large numerical constant $c > 0$.
3: **for** $\ell = h, \dots, 1$ **do**
4: $\quad\mathcal{D}_\ell \leftarrow \varnothing$.
5: $\quad$**for** n_{psdp} times **do**
6: $\quad\quad$Sample $\pi \sim p_\ell$.
7: $\quad\quad$Sample $(x_\ell, a_\ell, \sum_{\ell'=\ell}^h r_{\ell'}(x_{\ell'}, a_{\ell'})) \sim \pi \circ_\ell \pi_{\mathrm{unif}} \circ_{\ell+1} \widehat{\pi}$.
8: $\quad\quad$Update dataset: $\mathcal{D}_\ell \leftarrow \mathcal{D}_\ell \cup \left\{\left(x_\ell, a_\ell, \sum_{\ell'=\ell}^h r_{\ell'}(x_{\ell'}, a_{\ell'})\right)\right\}$.
9: $\quad$Solve regression:

$$\widehat{Q}_\ell \leftarrow \arg\min_{Q \in \mathcal{Q}_\ell} \sum_{(x,a,R)\in\mathcal{D}_\ell} (Q(x,a) - R)^2.$$

10: $\quad$Define $\widehat{\pi}_\ell(x) = \arg\max_{a\in\mathcal{A}} \widehat{Q}_\ell(x,a)$.
11: **return:** Policy $\widehat{\pi}$.

This section presents self-contained guarantees for Policy Search by Dynamic Programming (PSDP, Algorithm 6) (Bagnell et al., 2003), which performs local policy optimization given access to exploratory distributions $p_{1:h} \in \Delta(\Pi_{\mathsf{rns}})$. PSDP takes as input an arbitrary reward functions $r_{1:h} : \mathcal{X} \times \mathcal{A} \to [0,1]$ and a function class $\mathcal{Q} = \mathcal{Q}_{1:h}$, where $\mathcal{Q}_\ell \subseteq \{Q : \mathcal{X} \times \mathcal{A} \to [0,1]\}$, that can represent certain Q-functions for these rewards.

We prove that with high probability, the output $\widehat{\pi} = \mathsf{PSDP}_h(p_{1:h}, r_{1:h}; \epsilon, \delta, \mathcal{Q})$ is an approximate maximizer of the objective

$$\max_{\pi\in\Pi_{\mathsf{ns}}} \mathbb{E}^{M^\star,\pi}\left[\sum_{\ell=1}^h r_\ell(x_\ell, a_\ell)\right], \tag{58}$$

in a "local" sense with respect to $p_{1:h}$ (cf. Eq. (64)).

To prove guarantees for PSDP, we make use of the following realizability assumption for the class $\mathcal{Q} = \mathcal{Q}_{1:h}$.

Definition K.1. *We say that function classes $\mathcal{Q}_{1:h}$, where $\mathcal{Q}_\ell \subseteq \{Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+\}$ for $\ell \in [h]$, realize the reward functions $r_{1:h} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+$ if for all $t \in [h]$ and all $\pi \in \Pi_{\mathsf{ns}}$,*

$$Q_\ell^{M^\star,\pi}(\cdot,\cdot;r) \in \mathcal{Q}_\ell, \quad \textit{where} \quad Q_\ell^{M^\star,\pi}(x,a;r) := \mathbb{E}^{M^\star,\pi}\left[\sum_{\ell'=\ell}^h r_{\ell'}(x_{\ell'}, a_{\ell'}) \,\middle|\, x_\ell = x, a_\ell = a\right]. \tag{59}$$

Lemma K.5 (Main guarantee for PSDP)**.** *For any $\epsilon, \delta \in (0,1)$ and reward function $\{r_\ell\}_{\ell\in[h]}$ with $\sum_{\ell=1}^h r_\ell \in [0,1]$ that is realizable in the sense of Definition K.1, PSDP ensures that with probability at least $1-\delta$, the output $\widehat{\pi} = \mathsf{PSDP}_h(p_{1:h}, r_{1:h}; \epsilon, \delta, \mathcal{Q}_{1:h})$ satisfies*

$$\sum_{\ell=1}^{h} \mathbb{E}^{M^\star, p_\ell}\left[\max_{a\in\mathcal{A}} Q_\ell^{M^\star,\widehat{\pi}}(x_\ell, a; r) - Q_\ell^{M^\star,\widehat{\pi}}(x_\ell, \widehat{\pi}(x_\ell); r)\right] \le \epsilon, \tag{60}$$

and does so using at most $N_{\mathsf{psdp}}(\epsilon,\delta) = O\Big(\frac{H^3|\mathcal{A}|\log(|\mathcal{Q}|H\delta^{-1})}{\epsilon^2}\Big)$ episodes.

Proof of Lemma K.5. See the proof of Theorem D.1 in Mhammedi et al. (2023b). □

K.5. Proof of Theorem K.1

Theorem K.1 (General guarantee for Algorithm 3)**.** *Let $\varepsilon \in (0,1/2)$ and $\delta \in (0, e^{-1})$ be given, and suppose that Assumption K.1 and Assumption K.2 are satisfied. Then Algorithm 3 produces policy covers $p_1, \dots, p_H \in \Delta(\Pi_{\mathsf{rns}})$ such that with probability at least $1-\delta$, for all $h \in [H]$,*

$$\Psi^{M^\star}_{\mathsf{push};h,\varepsilon}(p_h) \le 170H\log(\varepsilon^{-1}) \cdot C^{M^\star}_{\mathsf{push}},$$

and does so using at most

$$N \le \widetilde{O}\Big(\frac{H|\mathcal{A}|\log(|\mathcal{W}|\delta^{-1})}{\varepsilon^3} + \frac{H}{\varepsilon} N_{\mathsf{opt}}(c\varepsilon^2, \delta/2HT)\Big)$$

episodes, where $c > 0$ is a sufficiently small absolute constant.

Proof of Theorem K.1. To keep notation compact, throughout this section we abbreviate $d_h^\pi \equiv d_h^{M^\star,\pi}$, $P_h(\cdot \mid \cdot) \equiv P_h^{M^\star}(\cdot \mid \cdot)$, $\mathbb{E}^\pi[\cdot] \equiv \mathbb{E}^{M,\pi}[\cdot]$, and so on when the MDP is clear from context.

Define

$$w_h^t(x_h \mid x_{h-1}, a_{h-1}) = \frac{P_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\sum_{i<t} d_h^{\pi^{h,i}}(x_h) + P_{h-1}(x_h \mid x_{h-1}, a_{h-1})} \tag{61}$$

We define two notions of estimation error for the weight function estimates produced by the subroutine EstimateWeightFunction (Algorithm 4):

$$(\varepsilon^t_{w,\mathrm{off};h})^2 = \mathbb{E}^{p_{h-1}}\left[\left(\sqrt{\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w_h^t(x_h \mid x_{h-1}, a_{h-1})}\right)^2\right], \quad \text{and} \tag{62}$$

$$(\varepsilon^t_{w,\mathrm{on};h})^2 = \frac{1}{t-1}\sum_{i<t} \mathbb{E}^{\pi^{h,i}\circ_{h-1}\pi_{\mathrm{unif}}}\left[\left(\sqrt{\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w_h^t(x_h \mid x_{h-1}, a_{h-1})}\right)^2\right]. \tag{63}$$

We define a notion of "local" suboptimality for the policies $\pi^{h,t}$ produced by the subroutine PolicyOptimization as follows:

$$\varepsilon^t_{\mathrm{opt};h} = \sum_{\ell=1}^{h-1} \mathbb{E}^{p_\ell}\left[\max_{a\in\mathcal{A}} Q_\ell^{\pi^{h,t}}(x_\ell, a; \widehat{w}_h^t) - Q_\ell^{\pi^{h,t}}(x_\ell, \pi^{h,t}(x_\ell); \widehat{w}_h^t)\right]. \tag{64}$$

Lemma K.6. *Let $\varepsilon \in (0,1/2)$ and set $T = \frac{1}{\varepsilon}$. Suppose that for all $h \ge 2$ and $t \in [T]$, it holds that $\varepsilon^t_{w,\mathrm{off};h} \le c_1(C^{M^\star}_{\mathsf{push}}/|\mathcal{A}|t)^{1/2}\varepsilon^{1/2}$, $\varepsilon^t_{w,\mathrm{on};h} \le c_2(C^{M^\star}_{\mathsf{push}}/|\mathcal{A}|t)^{1/2}\varepsilon^{1/2}$, and $\varepsilon^t_{\mathrm{opt};h} \le c_3\varepsilon^2$ for absolute constants $c_1, c_2, c_3 > 0$. Then for all $h \ge 2$, we have that*

$$\Psi^{M^\star}_{\mathsf{push};h,\varepsilon}(p_h) \le 170H\log(\varepsilon^{-1}) \cdot C^{M^\star}_{\mathsf{push}}.$$

Let $N_{\mathsf{weight}}(t,\epsilon,\delta)$ denote the number of episodes used by `EstimateWeightFunction` to ensure that $(\varepsilon^t_{w,\mathrm{off};h})^2, (\varepsilon^t_{w,\mathrm{on};h})^2 \le \epsilon^2/t$ with probability at least $1-\delta$ when invoked at iteration $t \in [T]$ for layer $h \ge 2$, and let $N_{\mathsf{opt}}(\epsilon,\delta)$ be the number of trajectories used by `PolicyOptimization` to ensures that $\varepsilon^t_{\mathrm{opt};h} \le \epsilon$ with probability at least $1-\delta$ when invoked at iteration $t \in [T]$ for layer $h \ge 2$. It follows from Lemma K.6 that with the parameter settings in Algorithm 3, we are guaranteed that with probability at least $1-\delta$, for all $h \in [H]$

$$\Psi^{M^\star}_{\mathsf{push};h,\varepsilon}(p_h) \le 170H\log(\varepsilon^{-1}) \cdot C^{M^\star}_{\mathsf{push}}.$$

and the total number of episodes used is at most

$$\begin{aligned} N &\le HT(N_{\mathsf{weight}}(T,\epsilon_w,\delta_w) + N_{\mathsf{opt}}(\epsilon_{\mathsf{opt}},\delta_{\mathsf{opt}})) \\ &\le HT\Big(N_{\mathsf{weight}}(T, c(C^{M^\star}_{\mathsf{push}}/|\mathcal{A}|)^{1/2}\varepsilon^{1/2}, \delta/2HT) + N_{\mathsf{opt}}(c'\varepsilon^2, \delta/2HT)\Big). \end{aligned}$$

for absolute constants $c, c' > 0$. It remains to bound $N_{\mathsf{weight}}(\epsilon,\delta)$, for which we appeal to the following lemma, a corollary of Theorem K.2.

Lemma K.7. *Let $h \ge 2$ and $t \in [T]$ be given. For any $\epsilon,\delta \in (0,1)$, distribution $p_{h-1} \in \Delta(\Pi_{\mathsf{rns}})$ and $\pi^{h,1},\dots,\pi^{h,t-1} \in \Pi_{\mathsf{ns}}$, `EstimateWeightFunction` ensures that with probability at least $1-\delta$, the output $\widehat{w}^t_h \leftarrow$ `EstimateWeightFunction`$_{h,t}(p_{h-1},\{\pi^{h,i}\}_{i<t};\epsilon,\delta,\mathcal{W})$ satisfies*

$$(\varepsilon^t_{w,\mathrm{off};h})^2 = \mathbb{E}^{M^\star,p_{h-1}}\left[\left(\sqrt{\widehat{w}^t_h(x_h \mid x_{h-1},a_{h-1})} - \sqrt{w^t_h(x_h \mid x_{h-1},a_{h-1})}\right)^2\right] \le \epsilon^2/t, \quad \textit{and} \tag{65}$$

$$(\varepsilon^t_{w,\mathrm{on};h})^2 = \frac{1}{t-1}\sum_{i<t}\mathbb{E}^{M^\star,\pi^{h,i}\circ_{h-1}\pi_{\mathrm{unif}}}\left[\left(\sqrt{\widehat{w}^t_h(x_h \mid x_{h-1},a_{h-1})} - \sqrt{w^t_h(x_h \mid x_{h-1},a_{h-1})}\right)^2\right] \le \epsilon^2/t, \tag{66}$$

and does so using at most $N_{\mathsf{weight}}(t,\epsilon,\delta) = 80t\frac{\log(|\mathcal{W}|\delta^{-1})}{\epsilon^2}$ episodes.

Appealing to this result, we conclude that the total number of episodes is at most

$$N \le \widetilde{O}\left(\frac{H|\mathcal{A}|\log(|\mathcal{W}|\delta^{-1})}{\varepsilon^3} + \frac{H}{\varepsilon}N_{\mathsf{opt}}(c\varepsilon^2,\delta/2HT)\right).$$

□

K.5.1. Proof of Lemma K.6 (Outer-Level Analysis)

Extended MDP. Let $\alpha \ge 1$ be a parameter to the proof, whose value will be chosen at the end as a function of ε. Following Mhammedi et al. (2023b;a), we define an *extended MDP* $\overline{M}$ by augmenting $\overline{\mathcal{A}} = \mathcal{A} \cup \{\mathfrak{t}\}$ and $\overline{\mathcal{X}} = \mathcal{X} \cup \{\mathfrak{t}\}$. $\overline{M}$ has identical dynamics to $M^\star$, except that taking action $\mathfrak{t}$ causes the state to transition to $\mathfrak{t}$ deterministically; $\mathfrak{t}$ is a self-looping terminal state.

We define $\overline{\Pi}_{\mathrm{RNS}}$ as the set of all randomized non-stationary policies from $\overline{\mathcal{X}}$ to $\overline{\mathcal{A}}$. For $\pi \in \overline{\Pi}_{\mathrm{RNS}}$, we abbreviate $\bar{d}^\pi_h \equiv d^{\overline{M},\pi}_h$, $\overline{P}_h(\cdot \mid \cdot) \equiv P^{\overline{M}}_h(\cdot \mid \cdot)$, $\overline{\mathbb{P}}^\pi[\cdot] \equiv \mathbb{P}^{\overline{M},\pi}[\cdot]$, $\overline{\mathbb{E}}^\pi[\cdot] \equiv \mathbb{E}^{\overline{M},\pi}[\cdot]$, and so on. We adopt the convention that all policies in $\overline{\Pi}_{\mathrm{RNS}}$ select action $\mathfrak{t}$ in the terminal state.

Truncated benchmark policy class. Given the policy class $\Pi = \Pi_{\mathsf{ns}}$, we inductively define a sequence of policy classes $\overline{\Pi}_{\alpha,1},\dots,\overline{\Pi}_{\alpha,H}$ based on the extended MDP and the output $p_1,\dots,p_H$ of the algorithm as follows.

- First, $\overline{\Pi}_{\alpha,0} = \Pi$.
- Next, for $h = 1,\dots,H$, we construct $\overline{\Pi}_{\alpha,h}$ from $\overline{\Pi}_{\alpha,h-1}$. For each $\pi \in \overline{\Pi}_{\alpha,h-1}$, add a policy π' to $\overline{\Pi}_{\alpha,h}$ defined as follows: For all $h' \ne h$, $\pi'_{h'} = \pi_{h'}$, and for layer h,

$$\pi'_h(x) = \begin{cases} \pi_h(x), & \frac{\bar{d}^\pi_h(x)}{\bar{d}^{p_h}_h(x)} \le \alpha, \\ \mathfrak{t}, & \frac{\bar{d}^\pi_h(x)}{\bar{d}^{p_h}_h(x)} > \alpha. \end{cases} \tag{67}$$

Finally, we adopt the shorthand $\overline{\Pi}_\alpha := \overline{\Pi}_{\alpha,H}$.

Main analysis. Let $h \geq 2$ be fixed. For the remainder of the proof, we abbreviate $\pi^t \equiv \pi^{t,h}$ to keep notation compact. Define

$$\overline{\Psi}_{\mathrm{push};h,\varepsilon}(p) = \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}\left[\frac{\overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\overline{d}_h^p(x_h) + \varepsilon \cdot \overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})} \mathbb{I}_{x_h \neq \mathfrak{t}}\right] \tag{68}$$

as a counterpart to the pushforward coverage relaxation for the extended MDP $\overline{M}$. We define three quantities,

$$\overline{\Delta}_{w,\mathrm{off};h} = \sum_{t=1}^{T} \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^{\pi}\left[\left(\sqrt{\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w_h^t(x_h \mid x_{h-1}, a_{h-1})}\right)^2 \mathbb{I}_{x_h \neq \mathfrak{t}}\right],$$

$$\overline{\Delta}_{w,\mathrm{on};h} = \sum_{t=1}^{T} \overline{\mathbb{E}}^{\pi^t}\left[\left(\sqrt{\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w_h^t(x_h \mid x_{h-1}, a_{h-1})}\right)^2 \mathbb{I}_{x_h \neq \mathfrak{t}}\right], \quad \text{and}$$

$$\overline{\Delta}_{\mathrm{opt};h} = \sum_{t=1}^{T} \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^{\pi}[\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1}) \mathbb{I}_{x_h \neq \mathfrak{t}}] - \overline{\mathbb{E}}^{\pi^t}[\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1}) \mathbb{I}_{x_h \neq \mathfrak{t}}].$$

which measure the quality of the weight function estimates produced by `EstimateWeightFunction` and the optimization quality of the policies produced by `PolicyOptimization` in the extended MDP.

We now state three technical lemmas, all proven in the sequel. The first lemma shows that we can bound the value of $\overline{\Psi}_{\mathrm{push};h,\varepsilon}(p_h)$ in terms of the pushforward coverability parameter $C^{M^\star}_{\mathsf{push};h}$, up to additive error terms depending on the quantities defined above.

Lemma K.8. *Let $h \geq 2$ be fixed and $\varepsilon \in (0, 1/2)$. Then, with $T = \frac{1}{\varepsilon}$, it holds that*

$$\overline{\Psi}_{\mathrm{push};h,\varepsilon}(p_h) \leq 72 C^{M^\star}_{\mathsf{push};h} \log(T) + 2\overline{\Delta}_{w,\mathrm{off};h} + 6\overline{\Delta}_{w,\mathrm{on};h} + 3\overline{\Delta}_{\mathrm{opt};h}. \tag{69}$$

The next two lemmas relate the weight estimation and policy optimization errors in $\overline{M}$ to their counterparts in the true MDP $M^\star$, leveraging key properties of the truncated policy class $\overline{\Pi}_\alpha$.

Lemma K.9. *The following bounds hold for all $h \geq 2$, as long as $\|w\|_\infty \leq 1$ for all $w \in \mathcal{W}_h$:*

$$\overline{\Delta}_{w,\mathrm{off};h} \leq \alpha|\mathcal{A}| \sum_{t=1}^{T} (\varepsilon^t_{w,\mathrm{off};h})^2, \quad \textit{and} \tag{70}$$

$$\overline{\Delta}_{w,\mathrm{on};h} \leq \sqrt{8|\mathcal{A}| C^{M^\star}_{\mathsf{push}} \log(T) \sum_{t=1}^{T} (t-1)(\varepsilon^t_{w,\mathrm{on};h})^2} + 4C^{M^\star}_{\mathsf{push}}. \tag{71}$$

Lemma K.10. *The following bound holds for all $h \geq 2$:*

$$\overline{\Delta}_{\mathrm{opt};h} \leq \alpha \sum_{t=1}^{T} \varepsilon^t_{\mathrm{opt};h}. \tag{72}$$

Appealing to Lemmas K.8 to K.10, we conclude that as long as $\varepsilon^t_{w,\mathrm{off};h} \leq c_1(\alpha|\mathcal{A}|t/C^{M^\star}_{\mathsf{push}})^{-1/2}$, $\varepsilon^t_{w,\mathrm{on};h} \leq c_2(|\mathcal{A}|Tt/C^{M^\star}_{\mathsf{push}})^{-1/2}$, and $\varepsilon^t_{\mathrm{opt};h} \leq c_3(\alpha T)^{-1}$ for all $h \geq 2$, $t \in [T]$, where $c_1, c_2, c_3 > 0$ are absolute constants, we are guaranteed that for all h,

$$\overline{\Psi}_{\mathrm{push};h,\varepsilon}(p_h) \leq 85 C^{M^\star}_{\mathsf{push}} \log(T). \tag{73}$$

It remains to translate this back to a bound on the L_1-Coverage objective for the true MDP $M^\star$. To do so, we start with the following technical lemma, also proven in the sequel.

Lemma K.11. *Consider any reward function $\{r_h\}_{h \in [H]}$ with $r_h : \overline{\mathcal{X}} \times \overline{\mathcal{A}} \to [0,1]$ such that $\sum_{h=1}^{H} r_h(x_h, a_h) \in [0,1]$ for all sequences $(x_1, a_1), \ldots, (x_H, a_H)$, and such that $r_h(\mathfrak{t}, a) = 0$ and $r_h(x, \mathfrak{t}) = 0$. It holds that*

$$\sup_{\pi \in \Pi} \mathbb{E}^{\pi}\left[\sum_{h=1}^{H} r_h\right] - \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^{\pi}\left[\sum_{h=1}^{H} r_h\right] \leq \sum_{h=1}^{H} \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{P}}^{\pi}\left[\frac{\overline{d}_h^{\pi}(x_h)}{\overline{d}_h^{p_h}(x_h)} > \alpha, x_h \neq \mathfrak{t}\right].$$

Let $h \geq 2$ be fixed and define a reward function $\{r_\ell\}_{\ell \leq h-1}$ with $r_\ell : \overline{\mathcal{X}} \times \overline{\mathcal{A}} \to [0, \varepsilon^{-1}]$ via

$$r_{h-1}(x,a) = \mathbb{E}\left[\frac{P_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{d_h^{p_h}(x_h) + \varepsilon \cdot P_{h-1}(x_h \mid x_{h-1}, a_{h-1})} \mid x_{h-1} = x, a_{h-1} = a\right]\mathbb{I}_{x,a \neq \mathfrak{t}}$$

and $r_\ell = 0$ for $\ell < h-1$. Using Lemma K.11, we have that

$$\Psi^{M^\star}_{\mathsf{push};h,\varepsilon}(p_h) = \sup_{\pi \in \Pi}\left[\sum_{\ell=1}^{h-1} r_\ell\right] \leq \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi\left[\sum_{\ell=1}^{h-1} r_\ell\right] + \frac{1}{\varepsilon}\sum_{\ell=1}^{h-1} \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{P}}^\pi\left[\frac{\bar{d}_\ell^\pi(x_\ell)}{\bar{d}_\ell^{p_\ell}(x_\ell)} > \alpha, x_\ell \neq \mathfrak{t}\right]. \tag{74}$$

To bound the right-hand side, we first note that

$$\begin{aligned}
\sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi\left[\sum_{\ell=1}^{h-1} r_\ell\right] &= \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi\left[\mathbb{E}\left[\frac{P_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{d_h^{p_h}(x_h) + \varepsilon \cdot P_{h-1}(x_h \mid x_{h-1}, a_{h-1})} \mid x_{h-1}, a_{h-1}\right]\mathbb{I}_{x_{h-1}, a_{h-1} \neq \mathfrak{t}}\right] \\
&= \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi\left[\overline{\mathbb{E}}\left[\frac{\overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\bar{d}_h^{p_h}(x_h) + \varepsilon \cdot \overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})} \mid x_{h-1}, a_{h-1}\right]\mathbb{I}_{x_{h-1}, a_{h-1} \neq \mathfrak{t}}\right] \\
&= \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi\left[\frac{\overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\bar{d}_h^{p_h}(x_h) + \varepsilon \cdot \overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\mathbb{I}_{x_{h-1}, a_{h-1} \neq \mathfrak{t}}\right] \\
&= \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi\left[\frac{\overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\bar{d}_h^{p_h}(x_h) + \varepsilon \cdot \overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\mathbb{I}_{x_h \neq \mathfrak{t}}\right] = \overline{\Psi}_{\mathrm{push};h,\varepsilon}(p_h).
\end{aligned} \tag{75}$$

Here, the second equality uses that i) $M^\star$ and $\overline{M}$ have identical transition dynamics whenever $x_{h-1}, a_{h-1}\mathfrak{t}$ and ii) policies in the support of p_h never take the terminal action. Meanwhile, the second-to-last inequality uses that $x_h \neq \mathfrak{t}$ if and only if $x_{h-1} \neq \mathfrak{t}$ and $a_{h-1} \neq \mathfrak{t}$ in $\overline{M}$. To bound the second term on the right-hand side of Eq. (74), we use a variant of Proposition 4.2.

Lemma K.12. *For all $\alpha > 0$ and $\ell \geq 1$, it holds that*

$$\sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{P}}^\pi\left[\frac{\bar{d}_\ell^\pi(x_\ell)}{\bar{d}_\ell^{p_\ell}(x_\ell)} > \alpha, x_\ell \neq \mathfrak{t}\right] \leq \frac{2}{\alpha} \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi\left[\frac{\overline{P}_{\ell-1}(x_\ell \mid x_{\ell-1}, a_{\ell-1})}{\bar{d}_\ell^{p_\ell}(x_\ell) + \alpha^{-1} \cdot \overline{P}_{\ell-1}(x_\ell \mid x_{\ell-1}, a_{\ell-1})}\mathbb{I}_{x_\ell \neq \mathfrak{t}}\right]. \tag{76}$$

We set $\alpha = \varepsilon^{-1}$, so that combining Eq. (74) with Eq. (75) and Lemma K.12 yields

$$\Psi^{M^\star}_{\mathsf{push};h,\varepsilon}(p_h) \leq \overline{\Psi}_{\mathrm{push};h,\varepsilon}(p_h) + 2\sum_{\ell=1}^{h-1} \overline{\Psi}_{\mathrm{push};\ell,\varepsilon}(p_\ell). \tag{77}$$

Consequently, for the choice $T = \varepsilon^{-1}$ and $\alpha = \varepsilon^{-1}$, Eq. (73) and Eq. (77) imply that for all $h \geq 2$,

$$\Psi^{M^\star}_{\mathsf{push};h,\varepsilon}(p_h) \leq 170H\log(\varepsilon^{-1}) \cdot C^{M^\star}_{\mathsf{push}}.$$

The final approximation requirements for these choices are $\varepsilon^t_{w,\mathrm{off};h} \leq c_1(C^{M^\star}_{\mathsf{push}}/|\mathcal{A}|t)^{1/2}\varepsilon^{1/2}$, $\varepsilon^t_{w,\mathrm{on};h} \leq c_2(C^{M^\star}_{\mathsf{push}}/|\mathcal{A}|t)^{1/2}\varepsilon^{1/2}$, and $\varepsilon^t_{\mathrm{opt};h} \leq c_3\varepsilon^2$ for absolute constants $c_1, c_2, c_3 > 0$. □

K.5.2. Proofs for Supporting Lemmas for Lemma K.6

Lemma K.8. *Let $h \geq 2$ be fixed and $\varepsilon \in (0, 1/2)$. Then, with $T = \frac{1}{\varepsilon}$, it holds that*

$$\overline{\Psi}_{\mathrm{push};h,\varepsilon}(p_h) \leq 72C^{M^\star}_{\mathsf{push};h}\log(T) + 2\overline{\Delta}_{w,\mathrm{off};h} + 6\overline{\Delta}_{w,\mathrm{on};h} + 3\overline{\Delta}_{\mathrm{opt};h}. \tag{69}$$

Proof of Lemma K.8. This is a slightly modified variant of the proof of Theorem 4.2. Let $\widetilde{d}_h^t = \sum_{i<t} d_h^{\pi^i}$ and $\check{d}_h^t = \sum_{i<t} \bar{d}_h^{\pi^i}$. Observe that for $T = \frac{1}{\varepsilon}$, we have

$$\begin{aligned}
\overline{\Psi}_{\mathrm{push};h,\varepsilon}(p_h) &= \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}\left[\frac{\overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\bar{d}_h^{p_h}(x_h) + \varepsilon \cdot \overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\mathbb{I}_{x_h \neq \mathfrak{t}}\right] \\
&= T \cdot \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi\left[\frac{\overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\check{d}_h^{T+1}(x_h) + \overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}\mathbb{I}_{x_h \neq \mathfrak{t}}\right],
\end{aligned}$$

and hence it suffices to bound the quantity on the right-hand side. Next, note that for all $t \in [T]$, we have that

$$\sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi \left[\frac{\overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\check{d}_h^{t+1}(x_h) + \overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})} \mathbb{I}_{x_h \neq \mathfrak{t}} \right] \leq \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi \left[\frac{\overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\check{d}_h^{t+1}(x_h) + \overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})} \mathbb{I}_{x_h \neq \mathfrak{t}} \right]$$

and consequently

$$T \cdot \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi \left[\frac{\overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\check{d}_h^{T+1}(x_h) + \overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})} \mathbb{I}_{x_h \neq \mathfrak{t}} \right] \leq \sum_{t=1}^T \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi \left[\frac{\overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\check{d}_h^{t}(x_h) + \overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})} \mathbb{I}_{x_h \neq \mathfrak{t}} \right].$$

Now, note that if $x_h \neq \mathfrak{t}$, the dynamics of $\overline{M}$ imply that we must have $x_{h-1}, a_{h-1} \neq \mathfrak{t}$ as well. In this case, we have

$$\frac{\overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\check{d}_h^{t}(x_h) + \overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})} = w_h^t(x_h \mid x_{h-1}, a_{h-1}), \tag{78}$$

since $\overline{P}_{h-1}(\cdot \mid x_{h-1}, a_{h-1}) = P_{h-1}(\cdot \mid x_{h-1}, a_{h-1})$ with $x_{h-1}, a_{h-1} \neq \mathfrak{t}$, and since $d_h^{\pi^i}(x_h) = \bar{d}_h^{\pi^i}(x_h)$ when $x_h \neq \mathfrak{t}$ (as the policies $\pi^1, \ldots, \pi^T$ never take the terminal action). As a result, using Lemma K.2, we have that

$$\begin{aligned}
&\sum_{t=1}^T \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi \left[\frac{\overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\check{d}_h^{t}(x_h) + \overline{P}_{h-1}(x_h \mid x_{h-1}, a_{h-1})} \mathbb{I}_{x_h \neq \mathfrak{t}} \right] \\
&= \sum_{t=1}^T \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi [w_h^t(x_h \mid x_{h-1}, a_{h-1}) \mathbb{I}_{x_h \neq \mathfrak{t}}] \\
&\leq 3 \sum_{t=1}^T \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi [\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1}) \mathbb{I}_{x_h \neq \mathfrak{t}}] \\
&\quad + 2 \underbrace{\sum_{t=1}^T \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi \left[\left(\sqrt{\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w_h^t(x_h \mid x_{h-1}, a_{h-1})} \right)^2 \mathbb{I}_{x_h \neq \mathfrak{t}} \right]}_{= \overline{\Delta}_{w,\mathrm{off};h}}.
\end{aligned}$$

Next, we can bound

$$\sum_{t=1}^T \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi [\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1}) \mathbb{I}_{x_h \neq \mathfrak{t}}] \leq \sum_{t=1}^T \overline{\mathbb{E}}^{\pi^t} [\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1}) \mathbb{I}_{x_h \neq \mathfrak{t}}] + \overline{\Delta}_{\mathrm{opt};h}$$

by definition. Applying Lemma K.2 once more, we have that

$$\begin{aligned}
&\sum_{t=1}^T \overline{\mathbb{E}}^{\pi^t} [\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1}) \mathbb{I}_{x_h \neq \mathfrak{t}}] \\
&\leq 3 \sum_{t=1}^T \overline{\mathbb{E}}^{\pi^t} [w_h^t(x_h \mid x_{h-1}, a_{h-1}) \mathbb{I}_{x_h \neq \mathfrak{t}}] + 2 \underbrace{\sum_{t=1}^T \overline{\mathbb{E}}^{\pi^t} \left[\left(\sqrt{\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w_h^t(x_h \mid x_{h-1}, a_{h-1})} \right)^2 \mathbb{I}_{x_h \neq \mathfrak{t}} \right]}_{= \overline{\Delta}_{w,\mathrm{on};h}}.
\end{aligned}$$

Finally, note that since $\pi^t \in \Pi$ never select the terminal action (and in particular never reach the terminal state), we have

$$\begin{aligned}
\sum_{t=1}^T \overline{\mathbb{E}}^{\pi^t} [w_h^t(x_h \mid x_{h-1}, a_{h-1}) \mathbb{I}_{x_h \neq \mathfrak{t}}] &= \sum_{t=1}^T \overline{\mathbb{E}}^{\pi^t} \left[\frac{P_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\widetilde{d}_h^t(x_h) + P_{h-1}(x_h \mid x_{h-1}, a_{h-1})} \mathbb{I}_{x_h \neq \mathfrak{t}} \right] \\
&= \sum_{t=1}^T \mathbb{E}^{\pi^t} \left[\frac{P_{h-1}(x_h \mid x_{h-1}, a_{h-1})}{\widetilde{d}_h^t(x_h) + P_{h-1}(x_h \mid x_{h-1}, a_{h-1})} \right] \\
&\leq 4 C_{\mathsf{push};h}^{M^\star} \log(2T),
\end{aligned}$$

where the final bound follows from is Eq. (35). To simplify the constants, we note that $\log(2T) \le 2\log(T)$ whenever $\varepsilon \le 1/2$.

□

Lemma K.9. *The following bounds hold for all $h \ge 2$, as long as $\|w\|_\infty \le 1$ for all $w \in \mathcal{W}_h$:*

$$\overline{\Delta}_{w,\mathrm{off};h} \le \alpha|\mathcal{A}| \sum_{t=1}^{T} (\varepsilon^t_{w,\mathrm{off};h})^2, \quad \textit{and} \tag{70}$$

$$\overline{\Delta}_{w,\mathrm{on};h} \le \sqrt{8|\mathcal{A}| C^{M^\star}_{\mathsf{push}} \log(T) \sum_{t=1}^{T} (t-1)(\varepsilon^t_{w,\mathrm{on};h})^2} + 4C^{M^\star}_{\mathsf{push}}. \tag{71}$$

Proof of Lemma K.9. We first bound the quantity

$$\overline{\Delta}_{w,\mathrm{off};h} = \sum_{t=1}^{T} \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^{\pi}\left[\left(\sqrt{\widehat{w}^t_h(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w^t_h(x_h \mid x_{h-1}, a_{h-1})}\right)^2 \mathbb{I}_{x_h \neq \mathfrak{t}}\right].$$

Observe that if $x_h \neq \mathfrak{t}$, then the dynamics of the MDP $\overline{M}$ imply that $x_{h-1} \neq \mathfrak{t}, a_{h-1} \neq \mathfrak{t}$. Consider an arbitrary policy $\pi \in \overline{\Pi}_\alpha$. Since $\pi(x_{h-1}) \neq \mathfrak{t}$, the dynamics in Eq. (67) imply that $\bar{d}^{\pi}_{h-1}(x_{h-1})/\bar{d}^{p_{h-1}}_{h-1}(x_{h-1}) \le \alpha$. Consequently, for any $t \in [T]$, we can bound

$$\begin{aligned}
&\overline{\mathbb{E}}^{\pi}\left[\left(\sqrt{\widehat{w}^t_h(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w^t_h(x_h \mid x_{h-1}, a_{h-1})}\right)^2 \mathbb{I}_{x_h \neq \mathfrak{t}}\right] \\
&= \sum_{x \in \mathcal{X}: \pi(x) \neq \mathfrak{t}, a \in \mathcal{A}} \bar{d}^{\pi}_{h-1}(x,a)\, \overline{\mathbb{E}}\left[\left(\sqrt{\widehat{w}^t_h(x_h \mid x, a)} - \sqrt{w^t_h(x_h \mid x, a)}\right)^2 \mid x_{h-1} = x, a_{h-1} = a\right] \\
&\le \sum_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}: \pi(x) \neq \mathfrak{t}} \bar{d}^{\pi}_{h-1}(x)\, \overline{\mathbb{E}}\left[\left(\sqrt{\widehat{w}^t_h(x_h \mid x, a)} - \sqrt{w^t_h(x_h \mid x, a)}\right)^2 \mid x_{h-1} = x, a_{h-1} = a\right] \\
&\le \alpha \sum_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \bar{d}^{p_{h-1}}_{h-1}(x)\, \overline{\mathbb{E}}\left[\left(\sqrt{\widehat{w}^t_h(x_h \mid x, a)} - \sqrt{w^t_h(x_h \mid x, a)}\right)^2 \mid x_{h-1} = x, a_{h-1} = a\right] \\
&= \alpha|\mathcal{A}| \cdot \overline{\mathbb{E}}^{p_{h-1} \circ_{h-1} \pi_{\mathrm{unif}}}\left[\left(\sqrt{\widehat{w}^t_h(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w^t_h(x_h \mid x_{h-1}, a_{h-1})}\right)^2\right] \\
&= \alpha|\mathcal{A}| \cdot \overline{\mathbb{E}}^{p_{h-1}}\left[\left(\sqrt{\widehat{w}^t_h(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w^t_h(x_h \mid x_{h-1}, a_{h-1})}\right)^2\right] \\
&= \alpha|\mathcal{A}| \cdot \mathbb{E}^{p_{h-1}}\left[\left(\sqrt{\widehat{w}^t_h(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w^t_h(x_h \mid x_{h-1}, a_{h-1})}\right)^2\right] = \alpha|\mathcal{A}| \cdot (\varepsilon^t_{w,\mathrm{off};h})^2,
\end{aligned}$$

where the second-to-last equality uses that $p_{h-1} \circ_{h-1} \pi_{\mathrm{unif}} = p_{h-1}$ by construction, and the final equality uses that policies in the support of p_{h-1} never take the optimal action. Summing over t completes the proof.

We now bound the quantity $\overline{\Delta}_{w,\mathrm{on};h}$. Note that since the policy π^t never chooses the terminal action $\mathfrak{t}$, we can write

$$\overline{\Delta}_{w,\mathrm{on};h} = \sum_{t=1}^{T} \mathbb{E}^{\pi^t}\left[\left(\sqrt{\widehat{w}^t_h(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w^t_h(x_h \mid x_{h-1}, a_{h-1})}\right)^2\right].$$

Observe that as a consequence of pushforward coverability, there exists a distribution $\mu \in \Delta(\mathcal{X})$ such that $d^{\pi}_{h-1}(x) \le C^{M^\star}_{\mathsf{push}} \mu(x)$ for all $x \in \mathcal{X}$, $\pi \in \Pi_{\mathsf{rns}}$. Hence, by applying Lemma K.3 with $g^t(x) =$

$\mathbb{E}\Big[\big(\sqrt{\widehat{w}_h^t(x_h \mid x, \pi^t(x))} - \sqrt{w_h^t(x_h \mid x, \pi^t(x))}\big)^2 \mid x_{h-1} = x, a_{h-1} = \pi^t(x)\Big]$, which has $g^t \in [0,2]$ whenever $\|\widehat{w}^t\|_\infty, \|w^t\|_\infty \le 1$, we have

$$
\begin{aligned}
&\overline{\Delta}_{w,\mathrm{on};h} \\
&\le \sqrt{2C_{\mathsf{push}}^{M^\star}\log(T)\sum_{\substack{t\in[T]\\ i<t}}\mathbb{E}^{\pi^i}\Big[\Big(\mathbb{E}\Big[\Big(\sqrt{\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w_h^t(x_h \mid x_{h-1}, a_{h-1})}\Big)^2 \mid x_{h-1}, a_{h-1} = \pi^t(x_{h-1})\Big]\Big)^2\Big]} \\
&\quad + 4C_{\mathsf{push}}^{M^\star} \\
&\le \sqrt{2C_{\mathsf{push}}^{M^\star}\log(T)\sum_{t=1}^{T}\sum_{i<t}\mathbb{E}^{\pi^i\circ_{h-1}\pi^t}\Big[\Big(\sqrt{\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w_h^t(x_h \mid x_{h-1}, a_{h-1})}\Big)^4\Big]} + 4C_{\mathsf{push}}^{M^\star} \\
&\le \sqrt{2C_{\mathsf{push}}^{M^\star}|\mathcal{A}|\log(T)\sum_{t=1}^{T}\sum_{i<t}\mathbb{E}^{\pi^i\circ_{h-1}\pi_{\mathrm{unif}}}\Big[\Big(\sqrt{\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w_h^t(x_h \mid x_{h-1}, a_{h-1})}\Big)^4\Big]} + 4C_{\mathsf{push}}^{M^\star} \\
&\le \sqrt{8C_{\mathsf{push}}^{M^\star}|\mathcal{A}|\log(T)\sum_{t=1}^{T}\sum_{i<t}\mathbb{E}^{\pi^i\circ_{h-1}\pi_{\mathrm{unif}}}\Big[\Big(\sqrt{\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w_h^t(x_h \mid x_{h-1}, a_{h-1})}\Big)^2\Big]} + 4C_{\mathsf{push}}^{M^\star}.
\end{aligned}
$$

This proves the result.

□

Lemma K.10. *The following bound holds for all $h \ge 2$:*

$$
\overline{\Delta}_{\mathrm{opt};h} \le \alpha \sum_{t=1}^{T} \varepsilon_{\mathrm{opt};h}^t. \tag{72}
$$

Proof of Lemma K.10. Let $t \in [T]$ be fixed, and recall that we can write

$$
\begin{aligned}
&\sup_{\pi\in\overline{\Pi}_\alpha} \overline{\mathbb{E}}^{\pi}[\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})\mathbb{I}_{x_h\neq\mathfrak{t}}] - \overline{\mathbb{E}}^{\pi^t}[\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})\mathbb{I}_{x_h\neq\mathfrak{t}}] \\
&= \sup_{\pi\in\overline{\Pi}_\alpha} \overline{\mathbb{E}}^{\pi}\big[\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})\mathbb{I}_{x_{h-1},a_{h-1}\neq\mathfrak{t}}\big] - \overline{\mathbb{E}}^{\pi^t}\big[\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})\mathbb{I}_{x_{h-1},a_{h-1}\neq\mathfrak{t}}\big],
\end{aligned}
$$

since $x_h \neq \mathfrak{t}$ if and only if $x_{h-1}, a_{h-1} \neq \mathfrak{t}$. If we define $\overline{Q}^\pi$ as the state-action value function for policy π in $\overline{M}$ under the reward function given by $r_{h-1}(x,a) = \overline{\mathbb{E}}[\widehat{w}_h^t(x_h \mid x, a) \mid x_{h-1} = x, a_{h-1} = a]\mathbb{I}_{x,a\neq\mathfrak{t}}$ and $r_{h'}(x,a) = 0$ for $h' < h-1$, the performance difference lemma (Kakade, 2003) implies that

$$
\begin{aligned}
&\sup_{\pi\in\overline{\Pi}_\alpha} \overline{\mathbb{E}}^{\pi}\big[\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})\mathbb{I}_{x_{h-1},a_{h-1}\neq\mathfrak{t}}\big] - \overline{\mathbb{E}}^{\pi^t}\big[\widehat{w}_h^t(x_h \mid x_{h-1}, a_{h-1})\mathbb{I}_{x_{h-1},a_{h-1}\neq\mathfrak{t}}\big] \\
&= \sup_{\pi\in\overline{\Pi}_\alpha} \sum_{\ell=1}^{h-1} \overline{\mathbb{E}}^{\pi}\Big[\overline{Q}_\ell^{\pi^t}(x_\ell, \pi(x_\ell)) - \overline{Q}_\ell^{\pi^t}(x_\ell, \pi^t(x_\ell))\Big].
\end{aligned}
$$

Consider an arbitrary policy $\pi \in \overline{\Pi}_\alpha$, and fix $\ell \in [h-1]$, and write

$$
\overline{\mathbb{E}}^{\pi}\Big[\overline{Q}_\ell^{\pi^t}(x_\ell, \pi(x_\ell)) - \overline{Q}_\ell^{\pi^t}(x_\ell, \pi^t(x_\ell))\Big] = \sum_{x\in\mathcal{X}} \bar{d}_\ell^{\pi}(x)\Big(\overline{Q}_\ell^{\pi^t}(x, \pi(x)) - \overline{Q}_\ell^{\pi^t}(x, \pi^t(x))\Big)
$$

For any $x \in \mathcal{X}$, since $\pi \in \overline{\Pi}_\alpha$, if $\bar{d}_\ell^{\pi}(x)/\bar{d}_\ell^{p_\ell}(x) > \alpha$, then $\pi(x) = \mathfrak{t}$, which implies that

$$
\overline{Q}_\ell^{\pi^t}(x, \pi(x)) - \overline{Q}_\ell^{\pi^t}(x, \pi^t(x)) \le -\overline{Q}_\ell^{\pi^t}(x, \pi^t(x)) \le 0
$$

since the reward is non-negative, and since we can receive non-zero reward only if $\pi(x_\ell) \neq \mathfrak{t}$ for all $\ell \leq h-1$. Hence, we can bound

$$\begin{aligned}
\sum_{x\in\mathcal{X}} \bar{d}^{\pi}_{\ell}(x)\Big(\overline{Q}^{\pi^t}_{\ell}(x,\pi(x)) - \overline{Q}^{\pi^t}_{\ell}(x,\pi^t(x))\Big) &\leq \sum_{x\in\mathcal{X}} \bar{d}^{\pi}_{\ell}(x)\Big(\overline{Q}^{\pi^t}_{\ell}(x,\pi(x)) - \overline{Q}^{\pi^t}_{\ell}(x,\pi^t(x))\Big)\mathbb{I}\big\{\bar{d}^{\pi}_{\ell}(x)/\bar{d}^{p_\ell}_{\ell}(x) \leq \alpha\big\} \\
&\leq \sum_{x\in\mathcal{X}} \bar{d}^{\pi}_{\ell}(x)\Big(\max_{a\in\mathcal{A}} \overline{Q}^{\pi^t}_{\ell}(x,a) - \overline{Q}^{\pi^t}_{\ell}(x,\pi^t(x))\Big)\mathbb{I}\big\{\bar{d}^{\pi}_{\ell}(x)/\bar{d}^{p_\ell}_{\ell}(x) \leq \alpha\big\} \\
&\leq \alpha \sum_{x\in\mathcal{X}} \bar{d}^{p_\ell}_{\ell}(x)\Big(\max_{a\in\mathcal{A}} \overline{Q}^{\pi^t}_{\ell}(x,a) - \overline{Q}^{\pi^t}_{\ell}(x,\pi^t(x))\Big)\mathbb{I}\big\{\bar{d}^{\pi}_{\ell}(x)/\bar{d}^{p_\ell}_{\ell}(x) \leq \alpha\big\} \\
&\leq \alpha \sum_{x\in\mathcal{X}} \bar{d}^{p_\ell}_{\ell}(x)\Big(\max_{a\in\mathcal{A}} \overline{Q}^{\pi^t}_{\ell}(x,a) - \overline{Q}^{\pi^t}_{\ell}(x,\pi^t(x))\Big) \\
&= \alpha\, \overline{\mathbb{E}}^{p_\ell}\Big[\max_{a\in\mathcal{A}} \overline{Q}^{\pi^t}_{\ell}(x,a) - \overline{Q}^{\pi^t}_{\ell}(x,\pi^t(x))\Big].
\end{aligned}$$

Above, the second inequality uses that i) $\overline{Q}^{\pi^t}_{\ell}(x,a) = 0$ for all $a \in \bar{\mathcal{A}}$ if $x = \mathfrak{t}$ and ii) $\pi(x) \in \mathcal{A}$ if $x \neq \mathfrak{t}$ but $\bar{d}^{\pi}_{\ell}(x)/\bar{d}^{p_\ell}_{\ell}(x) \leq \alpha$. Finally, we note that

$$\overline{\mathbb{E}}^{p_\ell}\Big[\max_{a\in\mathcal{A}} \overline{Q}^{\pi^t}_{\ell}(x_\ell,a) - \overline{Q}^{\pi^t}_{\ell}(x_\ell,\pi^t(x_\ell))\Big] = \mathbb{E}^{p_\ell}\Big[\max_{a\in\mathcal{A}} Q^{\pi^t}_{\ell}(x_\ell,a;\widehat{w}^t_h) - \overline{Q}^{\pi^t}_{\ell}(x_\ell,\pi^t(x_\ell);\widehat{w}^t_h)\Big],$$

since i) policies in the support of p_ℓ never take the terminal action, and ii) $\overline{Q}^{\pi^t}_{\ell}(x,a) = Q^{\pi^t}_{\ell}(x,a;\widehat{w}^t_h)$ whenever $x, a \neq \mathfrak{t}$, since π^t never takes the terminal action. □

Lemma K.11. *Consider any reward function $\{r_h\}_{h\in[H]}$ with $r_h : \overline{\mathcal{X}} \times \bar{\mathcal{A}} \to [0,1]$ such that $\sum_{h=1}^H r_h(x_h,a_h) \in [0,1]$ for all sequences $(x_1,a_1),\ldots,(x_H,a_H)$, and such that $r_h(\mathfrak{t},a) = 0$ and $r_h(x,\mathfrak{t}) = 0$. It holds that*

$$\sup_{\pi\in\Pi} \mathbb{E}^{\pi}\Big[\sum_{h=1}^H r_h\Big] - \sup_{\pi\in\overline{\Pi}_\alpha} \overline{\mathbb{E}}^{\pi}\Big[\sum_{h=1}^H r_h\Big] \leq \sum_{h=1}^H \sup_{\pi\in\overline{\Pi}_\alpha} \overline{\mathbb{P}}^{\pi}\Big[\frac{\bar{d}^{\pi}_h(x_h)}{\bar{d}^{p_h}_h(x_h)} > \alpha, x_h \neq \mathfrak{t}\Big].$$

Proof of Lemma K.11. Since policies $\pi \in \Pi$ never take the terminal action, we can write

$$\begin{aligned}
&\sup_{\pi\in\Pi} \mathbb{E}^{\pi}\Big[\sum_{h=1}^H r_h\Big] - \sup_{\pi\in\overline{\Pi}_\alpha} \overline{\mathbb{E}}^{\pi}\Big[\sum_{h=1}^H r_h\Big] \\
&= \sup_{\pi\in\overline{\Pi}_{\alpha,0}} \overline{\mathbb{E}}^{\pi}\Big[\sum_{h=1}^H r_h\Big] - \sup_{\pi\in\overline{\Pi}_{\alpha,H}} \overline{\mathbb{E}}^{\pi}\Big[\sum_{h=1}^H r_h\Big] \\
&= \sum_{\ell=1}^H \Bigg(\sup_{\pi\in\overline{\Pi}_{\alpha,\ell-1}} \overline{\mathbb{E}}^{\pi}\Big[\sum_{h=1}^H r_h\Big] - \sup_{\pi\in\overline{\Pi}_{\alpha,\ell}} \overline{\mathbb{E}}^{\pi}\Big[\sum_{h=1}^H r_h\Big]\Bigg)
\end{aligned}$$

by telescoping. Fix $\ell \in [H]$. Let $\pi \in \overline{\Pi}_{\alpha,\ell-1}$ be arbitrary, and let $\pi' \in \overline{\Pi}_{\alpha,\ell}$ denote the policy with $\pi'_h = \pi_h$ for $h \neq \ell$, and with

$$\pi'_\ell(x) = \begin{cases} \pi_\ell(x), & \frac{\bar{d}^{\pi}_\ell(x)}{\bar{d}^{p_\ell}_\ell(x)} \leq \alpha, \\ \mathfrak{t}, & \frac{\bar{d}^{\pi}_\ell(x)}{\bar{d}^{p_\ell}_\ell(x)} > \alpha. \end{cases} \tag{79}$$

Let $\overline{Q}^{\pi}_h(x,a) := \overline{\mathbb{E}}^{\pi}\Big[\sum_{h'=h}^H r_{h'} \mid x_h = x, a_h = a\Big]$ denote the Q-function for $\{r_h\}$ in $\overline{M}$. Then by the performance

difference lemma (Kakade, 2003), we have

$$\begin{aligned}
&\overline{\mathbb{E}}^{\pi}\left[\sum_{h=1}^{H} r_h\right] - \overline{\mathbb{E}}^{\pi'}\left[\sum_{h=1}^{H} r_h\right] \\
&= \overline{\mathbb{E}}^{\pi}\left[\sum_{h=1}^{H} \overline{Q}_h^{\pi'}(x_h, \pi(x_h)) - \overline{Q}_h^{\pi'}(x_h, \pi'(x_h))\right] \\
&= \overline{\mathbb{E}}^{\pi}\left[\overline{Q}_\ell^{\pi'}(x_\ell, \pi(x_\ell)) - \overline{Q}_\ell^{\pi'}(x_\ell, \pi'(x_\ell))\right],
\end{aligned}$$

since the policies agree unless $h = \ell$. Since $\overline{Q}_\ell^{\pi'} \in [0, 1]$ by the normalization assumption on the rewards and $\overline{Q}_\ell^{\pi'}(\mathfrak{t}, a) = 0$ for all $a \in \overline{\mathcal{A}}$, we have

$$\begin{aligned}
\overline{\mathbb{E}}^{\pi}\left[\overline{Q}_\ell^{\pi'}(x_\ell, \pi(x_\ell)) - \overline{Q}_\ell^{\pi'}(x_\ell, \pi'(x_\ell))\right] &\le \overline{\mathbb{P}}^{\pi}[\pi'(x_\ell) \neq \pi(x_\ell), x_\ell \neq \mathfrak{t}] \\
&= \overline{\mathbb{P}}^{\pi}\left[\frac{\bar{d}_\ell^{\pi}(x_\ell)}{\bar{d}_\ell^{p_\ell}(x_\ell)} > \alpha, x_\ell \neq \mathfrak{t}\right],
\end{aligned}$$

where the final equality follows from Eq. (79). Since this result holds uniformly for all $\pi \in \overline{\Pi}_{\alpha,\ell-1}$, we conclude that

$$\begin{aligned}
&\sup_{\pi \in \overline{\Pi}_{\alpha,\ell-1}} \overline{\mathbb{E}}^{\pi}\left[\sum_{h=1}^{H} r_h\right] - \sup_{\pi \in \overline{\Pi}_{\alpha,\ell}} \overline{\mathbb{E}}^{\pi}\left[\sum_{h=1}^{H} r_h\right] \\
&\le \sup_{\pi \in \overline{\Pi}_{\alpha,\ell-1}} \overline{\mathbb{P}}^{\pi}\left[\frac{\bar{d}_\ell^{\pi}(x_\ell)}{\bar{d}_\ell^{p_\ell}(x_\ell)} > \alpha, x_\ell \neq \mathfrak{t}\right] \\
&= \sup_{\pi \in \overline{\Pi}_{\alpha}} \overline{\mathbb{P}}^{\pi}\left[\frac{\bar{d}_\ell^{\pi}(x_\ell)}{\bar{d}_\ell^{p_\ell}(x_\ell)} > \alpha, x_\ell \neq \mathfrak{t}\right],
\end{aligned}$$

where the final equality uses that every policy in $\overline{\Pi}_{\alpha,\ell-1}$ has a counterpart in $\overline{\Pi}_\alpha = \overline{\Pi}_{\alpha,H}$ that takes identical actions for layers $1, \ldots, \ell - 1$.

□

Lemma K.12. *For all $\alpha > 0$ and $\ell \ge 1$, it holds that*

$$\sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{P}}^{\pi}\left[\frac{\bar{d}_\ell^{\pi}(x_\ell)}{\bar{d}_\ell^{p_\ell}(x_\ell)} > \alpha, x_\ell \neq \mathfrak{t}\right] \le \frac{2}{\alpha} \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^{\pi}\left[\frac{\overline{P}_{\ell-1}(x_\ell \mid x_{\ell-1}, a_{\ell-1})}{\bar{d}_\ell^{p_\ell}(x_\ell) + \alpha^{-1} \cdot \overline{P}_{\ell-1}(x_\ell \mid x_{\ell-1}, a_{\ell-1})} \mathbb{I}_{x_\ell \neq \mathfrak{t}}\right]. \tag{76}$$

Proof of Lemma K.12. We follow a proof similar to Proposition 4.2. For any ℓ, we can write

$$\begin{aligned}
&\sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{P}}^{\pi}\left[\frac{\bar{d}_\ell^{\pi}(x_\ell)}{\bar{d}_\ell^{p_\ell}(x_\ell)} > \alpha, x_\ell \neq \mathfrak{t}\right] \\
&= \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{P}}^{\pi}\left[\frac{2\bar{d}_\ell^{\pi}(x_\ell)}{\bar{d}_\ell^{p_\ell}(x_\ell) + \alpha^{-1}\bar{d}_\ell^{\pi}(x_\ell)} > \alpha, x_\ell \neq \mathfrak{t}\right] \\
&\le \frac{2}{\alpha} \sup_{\pi \in \overline{\Pi}_\alpha} \overline{\mathbb{E}}^{\pi}\left[\frac{\bar{d}_\ell^{\pi}(x_\ell)}{\bar{d}_\ell^{p_\ell}(x_\ell) + \alpha^{-1}\bar{d}_\ell^{\pi}(x_\ell)} \mathbb{I}_{x_\ell \neq \mathfrak{t}}\right] = \frac{2}{\alpha} \sup_{\pi \in \overline{\Pi}_\alpha} \sum_{x \in \overline{\mathcal{X}}} \frac{(\bar{d}_\ell^{\pi}(x))^2}{\bar{d}_\ell^{p_\ell}(x) + \alpha^{-1}\bar{d}_\ell^{\pi}(x)} \mathbb{I}_{x \neq \mathfrak{t}}.
\end{aligned}$$

By Lemma H.1, the function

$$d \mapsto \frac{(d)^2}{\bar{d}_\ell^{p_\ell}(x) + \alpha^{-1} \cdot d}$$

is convex for all x. Hence, writing $\bar{d}_\ell^{\pi}(x)\mathbb{I}_{x \neq \mathfrak{t}} = \overline{\mathbb{E}}^{\pi}\left[\overline{P}_{\ell-1}(x \mid x_{\ell-1}, a_{\ell-1})\mathbb{I}_{x \neq \mathfrak{t}}\right]$, Jensen's inequality implies that for all x,

$$\frac{(\bar{d}_\ell^{\pi}(x))^2}{\bar{d}_\ell^{p_\ell}(x) + \alpha^{-1} \cdot \bar{d}_\ell^{\pi}(x)} \mathbb{I}_{x \neq \mathfrak{t}} \le \overline{\mathbb{E}}^{\pi}\left[\frac{(\overline{P}_{\ell-1}(x \mid x_{\ell-1}, a_{\ell-1}))^2}{\bar{d}_\ell^{p_\ell}(x) + \alpha^{-1} \cdot \overline{P}_{\ell-1}(x \mid x_{\ell-1}, a_{\ell-1})} \mathbb{I}_{x \neq \mathfrak{t}}\right].$$

We conclude that

$$\sup_{\pi\in\overline{\Pi}_\alpha} \overline{\mathbb{P}}^\pi\left[\frac{\bar{d}^\pi_\ell(x_\ell)}{\bar{d}^{p_\ell}_\ell(x_\ell)} > \alpha, x_\ell \neq \mathfrak{t}\right] \leq \frac{2}{\alpha} \sup_{\pi\in\overline{\Pi}_\alpha} \overline{\mathbb{E}}^\pi\left[\frac{\overline{P}_{\ell-1}(x_\ell \mid x_{\ell-1}, a_{\ell-1})}{\bar{d}^{p_\ell}_\ell(x_\ell) + \alpha^{-1}\cdot \overline{P}_{\ell-1}(x_\ell \mid x_{\ell-1}, a_{\ell-1})}\mathbb{I}_{x_\ell\neq\mathfrak{t}}\right].$$

□

K.5.3. Proof of Lemma K.7 (Guarantee for `EstimateWeightFunction`)

Let $\overline{w}^t_h := t\cdot w^t_h$ and let $\check{w}^t_h := t\cdot \widehat{w}^t_h$. Observe that solving the optimization problem in Line 10 of Algorithm 4 is equivalent to solving the optimization problem in Eq. (54) over the class $t\cdot\mathcal{W}_h$, which has $\|w'\|_\infty \leq t$ for all $w'\in\mathcal{W}_h$. As such, we can appeal to Theorem K.2 (in particular, Remark K.2) with

$$\mu(x' \mid x, a) = P^{M^\star}_{h-1}(x' \mid x, a), \quad \nu(x' \mid x, a) = \frac{1}{t}\left(\sum_{i<t} d^{M^\star,\pi^{h,i}}_h(x') + P^{M^\star}_{h-1}(x' \mid x, a)\right),$$

and

$$\omega(x, a) = \frac{1}{2}\left(d^{M^\star,p_{h-1}}_{h-1}(x, a) + \frac{1}{t-1}\sum_{i<t} d^{M^\star,\pi^{h,i}\circ_{h-1}\pi_{\mathrm{unif}}}_{h-1}(x, a)\right).$$

Under Assumption K.1, we have

$$\frac{\mu(x' \mid x, a)}{\nu(x' \mid x, a)} = \overline{w}^t_h(x' \mid x, a) \in t\cdot\mathcal{W}_h,$$

so Theorem K.2 and Remark K.2 imply that

$$\mathbb{E}_{(x_{h-1},a_{h-1})\sim\omega}\left[\left(\sqrt{\check{w}^t_h(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{\overline{w}^t_h(x_h \mid x_{h-1}, a_{h-1})}\right)^2\right] \leq \frac{20t\log(|\mathcal{W}|\delta^{-1})}{tn} = \frac{20\log(|\mathcal{W}|\delta^{-1})}{n},$$

or equivalently,

$$\mathbb{E}_{(x_{h-1},a_{h-1})\sim\omega}\left[\left(\sqrt{\widehat{w}^t_h(x_h \mid x_{h-1}, a_{h-1})} - \sqrt{w^t_h(x_h \mid x_{h-1}, a_{h-1})}\right)^2\right] \leq \frac{20\log(|\mathcal{W}|\delta^{-1})}{tn}.$$

From the definition of ω, it follows that setting $n = n_{\mathsf{weight}}(\epsilon,\delta) = \frac{40\log(|\mathcal{W}|\delta^{-1})}{\epsilon^2}$ is sufficient to achieve the desired bound. The total number of episodes is at most $2t\cdot n_{\mathsf{weight}}(\epsilon,\delta)$.

□