# Minimax-Optimal Reward-Agnostic Exploration in Reinforcement Learning

Gen Li* Yuling Yan† Yuxin Chen* Jianqing Fan†

April 14, 2023

**Abstract**

This paper studies reward-agnostic exploration in reinforcement learning (RL) — a scenario where the learner is unware of the reward functions during the exploration stage — and designs an algorithm that improves over the state of the art. More precisely, consider a finite-horizon non-stationary Markov decision process with $S$ states, $A$ actions, and horizon length $H$, and suppose that there are no more than a polynomial number of given reward functions of interest. By collecting an order of

$$\frac{SAH^3}{\varepsilon^2} \text{ sample episodes} \quad \text{(up to log factor)}$$

without guidance of the reward information, our algorithm is able to find $\varepsilon$-optimal policies for all these reward functions, provided that $\varepsilon$ is sufficiently small. This forms the first reward-agnostic exploration scheme in this context that achieves provable minimax optimality. Furthermore, once the sample size exceeds $\frac{S^2AH^3}{\varepsilon^2}$ episodes (up to log factor), our algorithm is able to yield $\varepsilon$ accuracy for arbitrarily many reward functions (even when they are adversarially designed), a task commonly dubbed as "reward-free exploration." The novelty of our algorithm design draws on insights from offline RL: the exploration scheme attempts to maximize a critical reward-agnostic quantity that dictates the performance of offline RL, while the policy learning paradigm leverages ideas from sample-optimal offline RL paradigms.

**Keywords:** reward-agnostic exploration, reward-free exploration, offline reinforcement learning, sample complexity, minimax optimality

## Contents

**1 Introduction** **2**
- 1.1 Reward-agnostic exploration . . . 2
- 1.2 This paper . . . 3
- 1.3 Notation . . . 4

**2 Problem formulation** **4**

**3 Algorithm** **6**
- 3.1 A two-stage algorithm . . . 6
- 3.2 Subroutines: approximately solving the subproblems (14) and (20) . . . 8
- 3.3 Intuition . . . 11

**4 Main results** **13**

**5 Prior art** **14**

*Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA; Email: {ligen,yuxinc}@wharton.upenn.edu.

†Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; Email: {yulingy,jqfan}@princeton.edu.

**6 Analysis for reward-agnostic exploration (proof of Theorem 1) 15**
6.1 Preliminary facts . . . 15
6.2 Main steps . . . 16
6.3 Proof of Lemma 4 . . . 21

**7 Analysis for reward-free exploration 27**
7.1 Proof of Lemma 5 . . . 28

**8 Discussion 29**

**A A pessimistic model-based algorithm for offline RL 30**

**B Proof of Lemma 1 31**

# 1 Introduction

Understanding the efficacy and efficiency of exploration is one of the central research lines in statistical reinforcement learning (RL), dating back to earlier works like Lai and Robbins (1985) from the multi-armed bandit literature. In truth, a wealth of RL applications anticipates automated decision making without prior knowledge of the environment, for which exploration serves as an essential component to acquire and improve knowledge of the unknowns. With the aim to maximize information gain and enhance data efficiency, a large strand of recent works sought to balance the need to explore (i.e., gaining new information about under-explored states and actions) with the desire for exploitation (i.e., taking advantage of what is currently viewed to be favorable) in a statistically efficient manner (Auer and Ortner, 2006; Azar et al., 2017; Brafman and Tennenholtz, 2002; Jaksch et al., 2010; Jin et al., 2018; Kearns and Singh, 2002; Lattimore and Szepesvári, 2020; Li et al., 2022a; Simchowitz and Jamieson, 2019; Zanette and Brunskill, 2019; Zhang et al., 2020b).

## 1.1 Reward-agnostic exploration

Noteworthily, a dominant fraction of existing exploration schemes required prior information about the reward function, or at least information about those immediate rewards assocciated with the sampled state-action pairs. This requirement, however, becomes ill-suited for applications that do not come with pre-specified reward functions. For instance, in offline/batch RL, new learning tasks need to be performed on the basis of pre-collected data, but the reward function of this new task might not be readily available during the data pre-collection process; in online recommendation systems, when a recommendation is presented to a user, her feedback might not be instantaneously revealed, but instead be received in a considerably delayed manner. There is also no shortage of applications where the reward functions are frequently updated to meet multiple objectives or encourage different behavior, but it is clearly undesirable to recollect all data from scratch whenever the reward functions change. The breadth of these applications underscores the need to re-examine reward-agnostic exploration, namely,

> *How to explore the unknown environment in the most statistically efficient way even when the learner is unaware of the reward function(s) during the exploration stage? Can we perform exploration just once but still achieve efficiency for multiple unseen reward functions simultaneously?*

The design of reward-agnostic exploration or pure exploration is, as one would expect, substantially more challenging than the reward-aware counterpart. For the most part, existing reward-aware exploration schemes prioritize exploration of the "important" part of the environment in a task-specific manner, which could often be ill-suited to a new task with drastically different rewards. To address this issue, a natural remedy is to attempt exploration of all non-negligible states/actions thoroughly, in a way that is *simultaneously* adequate for all possible reward functions. This constitutes a key principle underlying a recent strand of works that goes by the name of *reward-free exploration (RFE)* (Jin et al., 2020a; Kaufmann et al., 2021; Zhang et al., 2021b). In the problem of reward-free exploration, one targets a pure exploration scheme such that the samples collected via this scheme are effective uniformly over all possible reward functions (including the ones that are adversarially designed). To facilitate more concrete discussion, imagine we are interested in a

| setting | paper | upper bound | number/type of reward functions |
|---|---|---|---|
| reward-agnostic | Zhang et al. (2020a) | $\frac{H^5SA}{\varepsilon^2}$ | $\mathsf{poly}(H,S,A)$, fixed |
| | this paper | $\frac{H^3SA}{\varepsilon^2}$ | |
| reward-free | Jin et al. (2020a) | $\frac{H^5S^2A}{\varepsilon^2}$ | arbitrary |
| | Kaufmann et al. (2021) | $\frac{H^4S^2A}{\varepsilon^2}$ | |
| | Ménard et al. (2021a) | $\frac{H^3S^2A}{\varepsilon^2}$ | |
| | Qiao et al. (2022) | $\frac{H^5S^2A}{\varepsilon^2}$ | |
| | this paper | $\frac{H^3S^2A}{\varepsilon^2}$ | |

Table 1: Sample complexity comparisons for reward-agnostic/reward-free algorithms on episodic non-stationary finite-horizon MDPs. For ease of presentation, we omit all logarithmic factors and lower-order terms.

finite-horizon non-stationary Markov decision process (MDP) with $S$ states, $A$ actions, and horizon length $H$. The state-of-the-art results on this front (Ménard et al., 2021a) asserted that: to enable reward-free exploration for an arbitrary set of reward functions, it suffices to sample[1]

$$\widetilde{O}\left(\frac{H^3S^2A}{\varepsilon^2}\right) \text{ episodes} \qquad \text{(reward-free exploration)} \tag{1}$$

during the exploration stage, where $\varepsilon$ denotes the target accuracy level. Notably, this bound is valid regardless of how many reward functions are under consideration.

The aim set out in the above RFE works, however, could sometimes be overly conservative in practice. There are plenty of practical scenarios where only a small number (or no more than a polynomial number) of reward functions matter. If this were the case, then the sample complexity bound (1) could have been too conservative. Following this motivation, a recent work Zhang et al. (2020a) came up with a model-free task-agnostic exploration algorithm that succeeds with

$$\widetilde{O}\left(\frac{H^5SA}{\varepsilon^2}\right) \text{ episodes}, \tag{2}$$

as long as the total number of reward functions is no larger than $\mathsf{poly}(S,A,H)$. In some aspect, this results in significant sample size saving if one onlys cares about a reasonable number of reward functions, given that the scaling with the number of states is dramatically improved (i.e., from $S^2$ to $S$) . Nevertheless, this bound (2) remains inferior to (1) — by a factor of $H^2$ — when it comes to the horizon dependency. A natural question arises that motivates the present paper: can we hope to develop a pure exploration scheme that is statistically optimal in terms of all salient parameters (i.e., $S, A, H$) of the MDP, provided that there are only a few fixed reward functions of interest?

## 1.2 This paper

One focus of this paper is to design a sample-efficient exploration algorithm, for a scenario where (i) there might be one or more fixed reward functions of interest, (ii) all reward functions are invisible to the learner during the exploration stage, and (iii) we are only allowed to collect data once but are asked to accommodate all these reward functions simultaneously. Given that prior research on reward-free exploration mostly

[1]For any two non-negative functions $f$ and $g$, the notation $f\big(S,A,H,\frac{1}{\varepsilon},\frac{1}{\delta}\big) = O\big(g\big(S,A,H,\frac{1}{\varepsilon},\frac{1}{\delta}\big)\big)$ means that there exists a universal constant $C_1 > 0$ such that $f\big(S,A,H,\frac{1}{\varepsilon},\frac{1}{\delta}\big) \le C_1 g\big(S,A,H,\frac{1}{\varepsilon},\frac{1}{\delta}\big)$. The notation $f\big(S,A,H,\frac{1}{\varepsilon},\frac{1}{\delta}\big) = \widetilde{O}\big(g\big(S,A,H,\frac{1}{\varepsilon},\frac{1}{\delta}\big)\big)$ is defined similarly except that it hides any logarithmic factor.

required reliable learning uniformly over all possible reward functions, the current paper adopts the terminology *reward-agnostic exploration* to differentiate it from RFE, allowing for the possibility that only a small number of reward functions count. To be more specific, the scenario studied herein consists of two stages: (a) the *exploration stage*: in which the learner explores the environment by collecting sample episodes in the absence of any reward information; (b) the *policy learning stage*: in which the learner is informed of the reward functions and computes a desired policy for each reward function of interest. Focusing on non-stationary finite-horizon MDPs as previously mentioned, our main contributions can be summarized below.

- *Reward-agnostic exploration.* Imagine that there are no more than $\mathsf{poly}(S, A, H)$ reward functions of interest. We put forward a reward-agnostic exploration scheme, which paired with a policy learning algorithm allows us to find an $\varepsilon$-optimal policy with a collection of

$$\widetilde{O}\left(\frac{H^3SA}{\varepsilon^2}\right) \text{ episodes}, \tag{3}$$

  ignoring the lower-order term. This sample complexity is essentially un-improvable — in a minimax sense — even when the reward function is *a priori* known (Domingues et al., 2021; Jin et al., 2018).

- *Reward-free exploration.* As it turns out, the proposed algorithm is also sample-efficient in the context of reward-free exploration; more precisely, it is able to return an $\varepsilon$-optimal policy uniformly over all sorts of reward functions (including those designed adversarially) with a sample size no larger than

$$\widetilde{O}\left(\frac{H^3S^2A}{\varepsilon^2}\right) \text{ episodes}. \tag{4}$$

  This sample complexity bound matches that of the best-performing RFE algorithms (i.e., Ménard et al. (2021a)) proposed in the literature.

More detailed comparisons between our results and past works are summarized in Table 1.

Before concluding, we remark that the crux of our approach is to leverage ideas from recent development in offline/batch RL, with the following key elements.

- In the exploration stage, we identify a sort of reward-irrelevant quantities — determined mainly by the occupancy distributions — that dictate the performance of offline policy learning. The exploration policy is then selected adaptively in an attempt to optimize such reward-irrelevant quantities.
- In the policy learning stage, we invoke a minimax optimal offline RL algorithm — namely, a pessimistic model-based approach (Li et al., 2022b) — that computes a near-optimal policy based on the samples collected in the exploration stage as well as the revealed reward information.

### 1.3 Notation

For any set $\mathcal{X}$, we denote by $\Delta(\mathcal{X})$ the probability simplex over $\mathcal{X}$. For any integer $m$, we define $[m] \coloneqq \{1, \ldots, m\}$. For any distribution $\rho \in \Delta(\mathcal{S})$ and any vector $V \in \mathbb{R}^S$, we define the associated variance as

$$\mathsf{Var}_\rho(V) \coloneqq \sum_{s\in\mathcal{S}} \rho(s)\big(V(s)\big)^2 - \Big(\sum_{s\in\mathcal{S}} \rho(s)V(s)\Big)^2. \tag{5}$$

For any two non-negative functions $f$ and $g$, the notation $f\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big) \lesssim g\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big)$ means that there exists a universal constant $C_1 > 0$ such that $f\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big) \le C_1 g\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big)$; the notation $f\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big) \gtrsim g\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big)$ indicates the existence of some universal constant $C_2 > 0$ such that $f\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big) \ge C_2 g\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big)$; and the notation $f\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big) \asymp g\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big)$ means that $f\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big) \lesssim g\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big)$ and $f\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big) \gtrsim g\big(S, A, H, \frac{1}{\varepsilon}, \frac{1}{\delta}\big)$ hold simultaneously.

## 2 Problem formulation

In this section, we introduce some preliminaries for Markov decision processes, and formulate the problem.

**Basics of Markov decision processes.** We study finite-horizon MDPs with horizon length $H$ and non-stationary probability transition kernels. The state space of the MDP is denoted by $\mathcal{S} = \{1, \ldots, S\}$, containing $S$ different states; the action space is denoted by $\mathcal{A} = \{1, \ldots, A\}$, comprising $A$ possible actions; we also let $P = \{P_h\}_{1 \le h \le H}$ (with $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$) represent the inhomogeneous probability transition kernel, such that taking action $a$ in state $s$ at step $h$ results in a new state at step $h+1$ following the distribution $P_h(\cdot \,|\, s, a)$. For notational convenience, we shall often adopt the shorthand notation

$$P_{h,s,a} \coloneqq P_h(\cdot \,|\, s, a) \in \Delta(\mathcal{S}). \tag{6}$$

A policy $\pi$ is a (possibly randomized) strategy for action selection. In particular, when policy $\pi$ is said to be Markovian (so that the action selection rule at step $h$ depends only on the state at this step), we often represent it as $\pi = \{\pi_h\}_{1 \le h \le H}$, where $\pi_h$ is a mapping from $\mathcal{S}$ to $\Delta(\mathcal{A})$, with $\pi_h(\cdot \,|\, s) \in \Delta(\mathcal{A})$ specifying how to choose actions at step $h$. When $\pi$ is a deterministic policy, we overload the notation by letting $\pi_h(s)$ denote the action chosen by $\pi$ in state $s$ at step $h$. Another ingredient that merits particular attention is the reward function, which does not affect how the MDP evolves. For any reward function $r = \{r_h\}_{1 \le h \le H}$ with $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$, the element $r_h(s, a)$ indicates the immediate reward gained in state $s$ and step $h$ while executing action $a$; here, we follow the convention by assuming $r_h(s, a) \in [0, 1]$ for each $(s, a, h)$ triple. The readers are also referred to standard textbooks like Bertsekas (2017) for more backgrounds.

For any policy $\pi$, the value function $V^\pi = \{V_h^\pi\}_{1 \le h \le H}$ (with $V_h^\pi : \mathcal{S} \to \mathbb{R}$) and the Q-function $Q^\pi = \{Q_h^\pi\}_{1 \le h \le H}$ (with $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$) associated with $\pi$ at step $h$ are defined and denoted by

$$\forall s \in \mathcal{S}: \qquad V_h^\pi(s) \coloneqq \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\middle|\, s_h = s; \pi\right], \tag{7a}$$

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \qquad Q_h^\pi(s, a) \coloneqq \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\middle|\, s_h = s, a_h = a; \pi\right], \tag{7b}$$

where the expectation is taken over the randomness of a sample trajectory $\{(s_t, a_t)\}_{t=h}^H$ induced by the underlying MDP under policy $\pi$. In words, the value function (resp. Q-function) quantifies the expected cumulative reward starting from step $h$ conditioned on a given state (resp. state-action pair) at step $h$. It is well-known that there exists at least one deterministic policy — denoted by $\pi^\star = \{\pi_h^\star\}_{1 \le h \le H}$ throughout — that simultaneously maximizes the value function and the Q-function for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Here and below, we shall refer to $\pi^\star$ as the optimal deterministic policy, and denote the optimal value function $V^\star = \{V_h^\star\}_{1 \le h \le H}$ and optimal Q-function $Q^\star = \{Q_h^\star\}_{1 \le h \le H}$ respectively as follows:

$$V_h^\star(s) = V_h^{\pi^\star}(s) = \sup_\pi V_h^\pi(s) \quad \text{and} \quad Q_h^\star(s, a) = Q_h^{\pi^\star}(s, a) = \sup_\pi Q_h^\pi(s, a), \quad \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{8}$$

When the initial state is drawn from a state distribution $\rho \in \Delta(\mathcal{S})$, we denote

$$V_1^\pi(\rho) \coloneqq \mathop{\mathbb{E}}_{s \sim \rho}\left[V_1^\pi(s)\right] \qquad \text{and} \qquad V_1^\star(\rho) \coloneqq \mathop{\mathbb{E}}_{s \sim \rho}\left[V_1^\star(s)\right]. \tag{9}$$

Additionally, we find it convenient to define the following occupancy distributions associated with policy $\pi$ at step $h$:

$$\forall s \in \mathcal{S}: \qquad d_h^\pi(s) \coloneqq \mathbb{P}\left(s_h = s \,|\, s_1 \sim \rho; \pi\right), \tag{10a}$$

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \qquad d_h^\pi(s, a) \coloneqq \mathbb{P}\left(s_h = s, a_h = a \,|\, s_1 \sim \rho; \pi\right), \tag{10b}$$

where the sample trajectory $\{(s_h, a_h)\}_{h=1}^H$ is generated according to the initial state distribution $s_1 \sim \rho$ and policy $\pi$. Evidently, the occupancy state distribution for step $h = 1$ reduces to

$$\forall s \in \mathcal{S}: \qquad d_1^\pi(s) = \rho(s). \tag{11}$$

**Learning processes and goals.** The learning process considered in this paper can be split into two separate stages. In the first stage — called the *exploration stage* — the learner executes the MDP sequentially

for $N_{\mathsf{tot}}$ times to collect $N_{\mathsf{tot}}$ sample episodes each of length $H$, using any exploration policies it selects. Throughout this paper, each sample episode starts from an initial state independently generated from a distribution $\rho \in \Delta(\mathcal{S})$, where $\rho$ is unknown to the learner. A key constraint, however, is that no reward information whatsoever can be utilized to guide data collection. In the second stage — termed the *policy learning stage* — the learner attempts to compute a policy on the basis of these $N_{\mathsf{tot}}$ episodes; no additional sampling is permitted during this stage.

The performance metric we focus on is the sample complexity; more precisely, given an initial state distribution $\rho$, a target accuracy level $\varepsilon$ and a reward function of interest, the sample complexity of an algorithm refers to the minimum number of sample episodes (during the exploration stage) that allows one to achieve $V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon$, with $\widehat{\pi}$ denoting the policy estimated by this algorithm. Depending on the number/type of reward functions of interest, we study the following two scenarios and use different names to distinguish them.

- *Reward-agnostic exploration.* In this case, we suppose that there exist $m_{\mathsf{reward}}$ given reward functions of interest, generated independently from the data samples. We are asked to compute a desirable policy for each of these reward functions, using the same set of data collected during the exploration stage. The aim is to achieve a sample complexity that scales optimally in $S, A, H$ and slowly in $m_{\mathsf{reward}}$.
- *Reward-free exploration.* This case assumes the presence of an arbitrarily large number $m_{\mathsf{reward}}$ of reward functions (e.g., $m_{\mathsf{reward}}$ could be (super)-exponential in $S$, $A$ and $H$ or even unbounded), where each reward function can even be adversarially chosen. We seek to develop an exploration and learning paradigm that achieves a desirable sample complexity independent from $m_{\mathsf{reward}}$, possibly at the price of larger scaling in other parameters like $S$. The statistical feasibility of this goal stems from a basic observation: if we have available enough samples to ensure accurate learning of the transition kernel, then one can find reliable policies regardless of what reward functions are in use.

# 3 Algorithm

In this section, we present the proposed procedure, followed by some intuitive explanation of the design rationale. Here and throughout, we denote

$$\Pi \;\coloneqq\; \text{the set of all deterministic policies.} \tag{12}$$

## 3.1 A two-stage algorithm

On a high level, the algorithm we propose consists of the following two stages:

- *Stage 1: reward-agnostic exploration.* In this stage, we take samples in the hope of exploring the unknown environment adequately. No reward information is available in this stage.
  - *Stage 1.1: estimating occupancy distributions.* For each step $h \in [H]$, we draw $N$ episodes of samples, which allow us to estimate the occupancy distributions at this step induced by any policy.
  - *Stage 1.2: computing a desirable behavior policy and drawing samples.* Equipped with our estimated occupancy distributions, we compute a desirable behavior policy $\widehat{\mu}_{\mathsf{b}}$ — taking the form of a finite mixture of deterministic policies — and employ it to draw $K$ episodes of samples.
- *Stage 2: policy learning via offline RL.* The reward functions are revealed in this stage, but no more samples can be obtained. Our algorithm then learns, for each reward function, a near-optimal policy by applying a model-based offline RL algorithm to the $K$ episodes drawn in Stage 1.2.

The total number of episodes of exploration in our algorithm is therefore $N_{\mathsf{tot}} = K + NH$, where all episodes are collected in Stage 1. In the sequel, we elaborate on the algorithm details, with the intuition explained in Section 3.3.

**Initialization.** Let us begin by describing how to initialize our algorithm. Generate $N$ sample episodes independently each of length 1. As mentioned previously, the initial states of these $N$ episodes are i.i.d. drawn from the distribution $\rho$, denoted by

$$s_1^{n,0} \overset{\text{i.i.d.}}{\sim} \rho, \qquad 1 \le n \le N.$$

For any policy $\pi$, we set the empirical occupancy distribution induced by $\pi$ for step $h = 1$ as follows:

$$\widehat{d}_1^{\pi}(s) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}(s_1^{n,0} = s), \qquad \forall s \in \mathcal{S}; \tag{13a}$$

$$\widehat{d}_1^{\pi}(s,a) = \widehat{d}_1^{\pi}(s)\pi_1(a \,|\, s), \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{13b}$$

**Stage 1.1: estimating occupancy distributions.** Armed with the estimates (13) for the initial step, we would like to continue estimating the occupancy distributions for the remaining steps. Towards this end, we proceed in a forward manner from $h = 1, \ldots, H-1$. Suppose we already have access to the empirical occupancy distribution $\widehat{d}_h^{\pi}$ w.r.t. every policy $\pi$ at step $h$. We then attempt estimation for the next step $h+1$ by acquiring a further set of samples, as described below.

- *Select an exploration policy.* Approximately solve the following convex program:

$$\widehat{\mu}^h \approx \arg\max_{\mu \in \Delta(\Pi)} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \log \left[ \frac{1}{KH} + \mathop{\mathbb{E}}_{\pi \sim \mu} \left[ \widehat{d}_h^{\pi}(s,a) \right] \right], \tag{14}$$

  whose rationale will be elucidated in Section 3.3. This leads to our exploration policy that assists in model estimation for step $h+1$:

$$\pi^{\mathsf{explore},h} = \mathbb{E}_{\pi \sim \widehat{\mu}^h}[\pi]. \tag{15}$$

  Here, $\pi^{\mathsf{explore},h}$ is a mixture of deterministic policies, with the weight vector $\widehat{\mu}^h$ chosen to solve an "infinite-dimensional" optimization problem in (14). Note, however, that instead of finding an exact solution (which could have an infinite support size and be computationally infeasible), we will introduce a tractable optimization-based subroutine in Section 3.2 to yield an approximate solution with finite support.

- *Sampling and estimation.* We collect $N$ independent episodes each of length $h+1$ — denoted by $\{s_1^{n,h}, a_1^{n,h}, s_2^{n,h}, a_2^{n,h}, \ldots, s_{h+1}^{n,h}\}_{1 \le n \le N}$ — using the exploration policy $\pi^{\mathsf{explore},h}$ (cf. (15)). Aggregate all sample transitions at the $h$-th step to construct $\widehat{P}_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ such that

$$\widehat{P}_h(s' \,|\, s,a) = \frac{\mathbb{1}(N_h(s,a) > \xi)}{\max\{N_h(s,a),\, 1\}} \sum_{n=1}^{N} \mathbb{1}(s_h^{n,h} = s, a_h^{n,h} = a, s_{h+1}^{n,h} = s') \tag{16}$$

  for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$; here, we take

$$N_h(s,a) = \sum_{n=1}^{N} \mathbb{1}\left(s_h^{n,h} = s, a_h^{n,h} = a\right), \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{17}$$

  and $\xi$ represents the following pre-specified quantity

$$\xi = c_\xi H^3 S^3 A^3 \log \frac{HSA}{\delta} \tag{18}$$

  for some large enough universal constant $c_\xi > 0$. In words, $\widehat{P}_h(\cdot \,|\, s,a)$ reflects the empirical transition frequency as long as there are sufficient samples visiting $(s,a)$ at step $h$; otherwise it is simply set to zero (so in this sense, $\widehat{P}_h$ is not a transition kernel itself). With the above empirical estimates (16) in place, we can compute, for any deterministic policy $\pi$, the following estimate of the empirical occupancy distribution induced by $\pi$ for step $h+1$:

$$\widehat{d}_{h+1}^{\pi}(s) = \left\langle \widehat{P}_h(s \,|\, \cdot,\cdot),\, \widehat{d}_h^{\pi}(\cdot,\cdot) \right\rangle, \qquad \forall s \in \mathcal{S}, \tag{19a}$$

$$\widehat{d}_{h+1}^{\pi}(s,a) = \widehat{d}_{h+1}^{\pi}(s)\pi_{h+1}(a \,|\, s), \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{19b}$$

**Stage 1.2: computing a behavior policy and drawing samples.** Armed with the expressions of the estimated occupancy distribution $\{\widehat{d}_h^{\pi}\}$ for any $\pi \in \Pi$, we propose to solve

$$\widehat{\mu}_{\mathsf{b}} \approx \arg\max_{\mu \in \Delta(\Pi)} \left\{ \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \log \left[ \frac{1}{KH} + \mathbb{E}_{\pi \sim \mu} \big[ \widehat{d}_h^{\pi}(s,a) \big] \right] \right\}. \tag{20}$$

which resembles (14) except that the objective function now involves summation over all $h$; as before, it will be approximately solved by means of a tractable subroutine (to be described momentarily). We then set

$$\widehat{\pi}_{\mathsf{b}} = \mathbb{E}_{\pi \sim \widehat{\mu}_{\mathsf{b}}}[\pi] \tag{21}$$

as a behavior policy, and sample $K$ independent sample trajectories each of length $H$ — denoted by $\big\{s_1^{n,\mathsf{b}}, a_1^{n,\mathsf{b}}, s_2^{n,\mathsf{b}}, a_2^{n,\mathsf{b}}, \ldots, s_H^{n,\mathsf{b}}\big\}_{1 \le n \le K}$.

**Stage 2: policy learning via offline RL.** With the above $K$ episodes $\big\{s_1^{n,\mathsf{b}}, a_1^{n,\mathsf{b}}, s_2^{n,\mathsf{b}}, a_2^{n,\mathsf{b}}, \ldots, s_H^{n,\mathsf{b}}\big\}_{1 \le n \le K}$ at hand, we compute the final policy estimate $\widehat{\pi}$ by running a sample-efficient offline RL algorithm. A candidate offline RL algorithm is a pessimistic model-based algorithm studied in Li et al. (2022b). There is a slight difference here: based on our empirical estimates of the occupancy distributions, we can readily calculate the following quantity:

$$\widehat{N}_h^{\mathsf{b}}(s,a) = \left[ \frac{K}{4} \mathop{\mathbb{E}}_{\pi \sim \widehat{\mu}_{\mathsf{b}}} \big[ \widehat{d}_h^{\pi}(s,a) \big] - \frac{K\xi}{8N} - 3 \log \frac{HSA}{\delta} \right]_+, \tag{22}$$

where $[x]_+ \coloneqq \max\{x, 0\}$. As we shall prove later in Section 6.2, $\widehat{N}_h^{\mathsf{b}}(s,a)$ serves as — with high probability — a lower bound on the total number of visits to $(s,a,h)$ among these $K$ sample episodes, which will be employed to subsample the sample transitions (instead of exploiting two-fold sample splitting as in Li et al. (2022b)) and construct lower confidence bounds. Precise descriptions of this offline RL algorithm can be found in Appendix A.

The whole procedure of the proposed algorithm is summarized in Algorithm 1.

## 3.2 Subroutines: approximately solving the subproblems (14) and (20)

Thus far, we have not yet specified how to approximately solve the subproblems (14) and (20). As it turns out, a simple Frank-Wolfe type procedure (Frank and Wolfe, 1956) allows one to solve these efficiently, to be detailed in this subsection.

**Subroutine for solving the subproblem (20).** Let us start by tackling the convex subproblem (20), whose objective function is given by

$$f(\mu) \coloneqq \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \log \left[ \frac{1}{KH} + \mathop{\mathbb{E}}_{\pi \sim \mu} \big[ \widehat{d}_h^{\pi}(s,a) \big] \right]. \tag{25}$$

As can be easily calculated, the first variation[2] of $f$ at a measure $\mu \in \Delta(\Pi)$ is given by

$$\delta f(\mu): \quad \pi \mapsto \underbrace{\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{\widehat{d}_h^{\pi}(s,a)}{\frac{1}{KH} + \mathbb{E}_{\pi' \sim \mu} \big[ \widehat{d}_h^{\pi'}(s,a) \big]}}_{=: \delta f(\mu)(\pi)} \tag{26}$$

[2]The first variation of $f: \Delta(\Pi) \to \mathbb{R}$ at $\mu$ is defined as any measurable function $\delta f(\mu): \Pi \to \mathbb{R}$ that satisfies

$$\lim_{\varepsilon \to 0} \frac{f(\mu + \varepsilon \mathcal{X}) - f(\mu)}{\varepsilon} = \int \delta f(\mu) \mathrm{d}\mathcal{X}$$

for any signed measure $\mathcal{X}$ over $\Pi$ satisfying $\int \mathrm{d}\mathcal{X} = 0$. The first variation is defined up to an additive constant. See Gelfand et al. (2000) for more details.

**Algorithm 1:** Reward-agnostic RL: main algorithm.

**Input:** state space $\mathcal{S}$, action space $\mathcal{A}$, horizon length $H$, initial state distribution $\rho$, target success probability $1-\delta$, threshold $\xi = c_\xi H^3 S^3 A^3 \log(HSA/\delta)$.

/* Stage 1: reward-agnostic exploration */

/* Stage 1.1: estimating occupancy distributions */

Draw $N$ i.i.d. initial states $s_1^{n,0} \overset{\text{i.i.d.}}{\sim} \rho$ $(1 \le n \le N)$, and define the following functions

$$\widehat{d}_1^{\pi}(s) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{s_1^{n,0} = s\}, \qquad \widehat{d}_1^{\pi}(s,a) = \widehat{d}_1^{\pi}(s)\pi_1(a\,|\,s) \tag{23}$$

for any deterministic policy $\pi : \mathcal{S}\times[H] \to \Delta(\mathcal{A})$ and any $(s,a)\in\mathcal{S}\times\mathcal{A}$. (Note that these functions are defined for future purpose and not computed for the moment, as we have not specified policy $\pi$.)

**for** $h = 1$ **to** $H-1$ **do**

  Call Algorithm 2 to compute an exploration policy $\pi^{\mathsf{explore},h}$.

  Draw $N$ independent trajectories $\{s_1^{n,h}, a_1^{n,h}, \ldots, s_{h+1}^{n,h}\}_{1\le n\le N}$ using policy $\pi^{\mathsf{explore},h}$ and compute

$$\widehat{P}_h(s'\,|\,s,a) = \frac{\mathbb{1}(N_h(s,a) > \xi)}{\max\{N_h(s,a),\,1\}} \sum_{n=1}^{N} \mathbb{1}(s_h^{n,h} = s, a_h^{n,h} = a, s_{h+1}^{n,h} = s'), \qquad \forall (s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S},$$

  where $N_h(s,a) = \sum_{n=1}^{N} \mathbb{1}\{s_h^{n,h} = s, a_h^{n,h} = a\}$.

  For any deterministic policy $\pi : \mathcal{S}\times[H]\to\Delta(\mathcal{A})$ and any $(s,a)\in\mathcal{S}\times\mathcal{A}$, define

$$\widehat{d}_{h+1}^{\pi}(s) = \big\langle \widehat{P}_h(s\,|\,\cdot,\cdot),\, \widehat{d}_h^{\pi}(\cdot,\cdot)\big\rangle, \qquad \widehat{d}_{h+1}^{\pi}(s,a) = \widehat{d}_{h+1}^{\pi}(s)\pi_{h+1}(a\,|\,s). \tag{24}$$

/* Stage 1.2: computing a behavior policy and drawing samples */

Call Algorithm 3 to compute a behavior policy $\pi_{\mathsf{b}}$.

Draw $K$ independent sample trajectories $\big\{s_1^{n,\mathsf{b}}, a_1^{n,\mathsf{b}}, s_2^{n,\mathsf{b}}, a_2^{n,\mathsf{b}}, \ldots, s_H^{n,\mathsf{b}}\big\}_{1\le n\le K}$ using policy $\pi_{\mathsf{b}}$.

/* Stage 2: policy learning via offline RL */

For each reward function of interest, call Algorithm 4 to compute a policy estimate $\widehat{\pi}$.

**Output:** the policy estimate $\widehat{\pi}$ for each reward function of interest.

for any deterministic policy $\pi\in\Pi$. Intuitively, when $\mu,\nu\in\Delta(\Pi)$ are close, the first variation allows one to approximate $f(\nu)$ with the first-order expansion $f(\mu) + \int \delta f(\mu)\mathrm{d}(\nu-\mu)$. We then propose to solve (20) via the Frank-Wolfe algorithm, where each iteration consists of the following two steps:

- (*direction finding*) Find a direction $y^{(t)}\in\Delta(\Pi)$ that is a solution to the following problem

$$\arg\max_{\mu\in\Delta(\Pi)} \int \delta f\big(\mu_{\mathsf{b}}^{(t)}\big)\mathrm{d}\mu. \tag{27}$$

  Given that $\delta f(\mu)(\pi)$ is always non-negative (cf. (26)), one can simply take

$$y^{(t)} = \mathbb{1}\left(\pi^{(t)}\right) \qquad \text{with } \pi^{(t)} \in \arg\max_{\pi\in\Pi}\ \delta f(\mu_{\mathsf{b}}^{(t)})(\pi). \tag{28}$$

  Here, $\mathbb{1}(\pi)$ is a Dirac measure centred on $\pi\in\Pi$. It then boils down to maximizing $\delta f(\mu)(\pi)$ over all deterministic policies $\pi$. While this might seem like a challenging task at first glance, one can solve it by applying dynamic programming to an MDP associated with $\widehat{P}$. Note, however, that $\widehat{P}$ is not yet a valid probability transition kernel due to the truncation operation, and hence needs to be slightly modified. More concretely,

  - Introduce a finite-horizon MDP $\mathcal{M}_{\mathsf{b}} = (\mathcal{S}\cup\{s_{\mathsf{aug}}\}, \mathcal{A}, H, \widehat{P}^{\mathsf{aug}}, r_{\mathsf{b}})$, where $s_{\mathsf{aug}}$ is an augmented state and the reward function is chosen to be

$$r_{\mathsf{b},h}(s,a) = \begin{cases} \dfrac{1}{\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu_{\mathsf{b}}^{(t)}}\left[\widehat{d}_h^{\pi}(s,a)\right]} \in [0, KH], & \text{if } (s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]; \\ 0, & \text{if } (s,a,h)\in\{s_{\mathsf{aug}}\}\times\mathcal{A}\times[H]. \end{cases} \tag{29}$$

**Algorithm 2:** Frank-Wolfe-type subroutine for solving (14) for step $h$.

**Initialize**: $\mu^{(0)} = \delta_{\pi_{\mathsf{init}}}$ for an arbitrary policy $\pi_{\mathsf{init}} \in \Pi$, $T_{\max} = \lfloor 50SA\log(KH) \rfloor$.

**for** $t = 0$ **to** $T_{\max}$ **do**

Compute the optimal deterministic policy $\pi^{(t),\mathsf{b}}$ of the MDP $\mathcal{M}_{\mathsf{b}}^h = (\mathcal{S} \cup \{s_{\mathsf{aug}}\}, \mathcal{A}, H, \widehat{P}^{\mathsf{aug},h}, r_{\mathsf{b}}^h)$, where $r_{\mathsf{b}}^h$ is defined in (35), and $\widehat{P}^{\mathsf{aug},h}$ is defined in (36); let $\pi^{(t)}$ be the corresponding optimal deterministic policy of $\pi^{(t),\mathsf{b}}$ in the original state space. `// find the optimal policy`

Compute `// choose the stepsize`

$$\alpha_t = \frac{\frac{1}{SA} g(\pi^{(t)}, \widehat{d}, \mu^{(t)}) - 1}{g(\pi^{(t)}, \widehat{d}, \mu^{(t)}) - 1}, \quad \text{where} \quad g(\pi, \widehat{d}, \mu) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\frac{1}{KH} + \widehat{d}_h^{\pi}(s,a)}{\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu}[\widehat{d}_h^{\pi}(s,a)]}.$$

Here, $\widehat{d}_h^{\pi}(s,a)$ is computed via (23) for $h = 1$, and (24) for $h \geq 2$.

If $g(\pi^{(t)}, \widehat{d}, \mu^{(t)}) \leq 2SA$ then exit for-loop. `// stopping rule`

Update `// Frank-Wolfe update`

$$\mu^{(t+1)} = (1 - \alpha_t)\, \mu^{(t)} + \alpha_t\, \mathbb{1}_{\pi^{(t)}}\,.$$

**Output:** the exploration policy $\pi^{\mathsf{explore},h} = \mathbb{E}_{\pi\sim\mu^{(t)}}[\pi]$ and the weight $\widehat{\mu}^h = \mu^{(t)}$.

In addition, the augmented probability transition kernel $\widehat{P}^{\mathsf{aug}}$ is constructed based on $\widehat{P}$ as follows:

$$\widehat{P}_h^{\mathsf{aug}}(s' \,|\, s, a) = \begin{cases} \widehat{P}_h(s' \,|\, s, a), & \text{if } s' \in \mathcal{S} \\ 1 - \sum_{s'\in\mathcal{S}} \widehat{P}_h(s' \,|\, s, a), & \text{if } s' = s_{\mathsf{aug}} \end{cases} \quad \text{for all } (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]; \tag{30a}$$

$$\widehat{P}_h^{\mathsf{aug}}(s' \,|\, s_{\mathsf{aug}}, a) = \mathbb{1}(s' = s_{\mathsf{aug}}) \quad \text{for all } (a,h) \in \mathcal{A}\times[H]. \tag{30b}$$

Clearly, the augmented state is an absorbing state associated with zero immediate rewards.

– As can be easily seen, maximizing $\frac{\partial f(\mu)}{\partial \mu(\pi)}$ (cf. (26)) can be relaxed to finding the optimal policy of $\mathcal{M}_{\mathsf{b}}$, which can be accomplished efficiently using classical dynamic programming methods (Bertsekas, 2017).

- *(update)* Set the iterate to be

$$\mu_{\mathsf{b}}^{(t+1)} = (1 - \alpha_t)\mu_{\mathsf{b}}^{(t)} + \alpha_t\, \mathbb{1}\left(\pi^{(t)}\right), \tag{31}$$

where $0 < \alpha_t < 1$ denotes the learning rate. The learning rate $\alpha_t$ is chosen to be

$$\alpha_t = \frac{\frac{1}{SAH} g\left(\pi^{(t)}, \widehat{d}, \mu_{\mathsf{b}}^{(t)}\right) - 1}{g\left(\pi^{(t)}, \widehat{d}, \mu_{\mathsf{b}}^{(t)}\right) - 1}, \tag{32}$$

where we define

$$g\left(\pi, \widehat{d}, \mu\right) \coloneqq \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\frac{1}{KH} + \widehat{d}_h^{\pi}(s,a)}{\frac{1}{KH} + \mathbb{E}_{\pi'\sim\mu}\left[\widehat{d}_h^{\pi'}(s,a)\right]}. \tag{33}$$

**Stopping rule and iteration complexity.** In order to achieve computational efficiency, we would like to terminate the algorithm after a reasonable number of iterations. Towards this end, we propose the following stopping rule: if the following condition is met:

$$g\left(\pi^{(t)}, \widehat{d}, \mu_{\mathsf{b}}^{(t)}\right) \leq 2HSA, \tag{34}$$

then the subroutine stops and returns $\mu_{\mathsf{b}}^{(t)}$. As it turns out, this stopping rule yields a reasonable iteration complexity, as asserted by the following lemma. The proof can be found in Section B.

**Lemma 1.** *Armed with the learning rate (32) and the stopping rule (34), this subroutine terminates within* $O\left(HSA\log(KH)\right)$ *iterations.*

**Algorithm 3:** Frank-Wolfe-type subroutine for solving (20).

**Initialize**: $\mu_{\mathsf{b}}^{(0)} = \delta_{\pi_{\mathsf{init}}}$ for an arbitrary policy $\pi_{\mathsf{init}} \in \Pi$, $T_{\max} = \lfloor 50SAH\log(KH) \rfloor$.

**for** $t = 0$ **to** $T_{\max}$ **do**

Compute the optimal deterministic policy $\pi^{(t),\mathsf{b}}$ of the MDP $\mathcal{M}_{\mathsf{b}} = (\mathcal{S} \cup \{s_{\mathsf{aug}}\}, \mathcal{A}, H, \widehat{P}^{\mathsf{aug}}, r_{\mathsf{b}})$, where $r_{\mathsf{b}}$ is defined in (29), and $\widehat{P}^{\mathsf{aug}}$ is defined in (30); let $\pi^{(t)}$ be the corresponding optimal deterministic policy of $\pi^{(t),\mathsf{b}}$ in the original state space. `// find the optimal policy`

Compute `// choose the stepsize`

$$\alpha_t = \frac{\frac{1}{SAH} g(\pi^{(t)}, \widehat{d}, \mu_{\mathsf{b}}^{(t)}) - 1}{g(\pi^{(t)}, \widehat{d}, \mu_{\mathsf{b}}^{(t)}) - 1}, \quad \text{where} \quad g(\pi, \widehat{d}, \mu) = \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\frac{1}{KH} + \widehat{d}_h^{\pi}(s,a)}{\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu}\big[\widehat{d}_h^{\pi}(s,a)\big]}.$$

Here, $\widehat{d}_h^{\pi}(s,a)$ is computed via (23) for $h = 1$, and (24) for $h \geq 2$.

If $g(\pi^{(t)}, \widehat{d}, \mu_{\mathsf{b}}^{(t)}) \leq 2HSA$ then exit for-loop. `// stopping rule`

Update `// Frank-Wolfe update`

$$\mu_{\mathsf{b}}^{(t+1)} = (1-\alpha_t)\,\mu_{\mathsf{b}}^{(t)} + \alpha_t\, \mathbb{1}_{\pi^{(t)}}\,.$$

**Output:** the behavior policy $\pi_{\mathsf{b}} = \mathbb{E}_{\pi\sim\mu_{\mathsf{b}}^{(t)}}[\pi]$ and the associated weight $\widehat{\mu}_{\mathsf{b}} = \mu_{\mathsf{b}}^{(t)}$.

**Subroutine for solving the subproblem (14).** Approximately solving the subproblem (14) can be accomplished by means of roughly the same procedure for solving (20). The main thing that needs to be slightly modified is the construction of the auxiliary MDP. More specifically, for solving (14) w.r.t. step $h$ in the $t$-th iteration, we construct $\mathcal{M}_{\mathsf{b}}^h = (\mathcal{S} \cup \{s^{\mathsf{aug}}\}, \mathcal{A}, H, \widehat{P}^{\mathsf{aug},h}, r_{\mathsf{b}}^h)$, where $s_{\mathsf{aug}}$ is an augmented state as before, and the reward function is chosen to be

$$r_{\mathsf{b},j}^h(s,a) = \begin{cases} \frac{1}{\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu^{(t)}}\left[\widehat{d}_h^{\pi}(s,a)\right]} \in [0, KH], & \text{if } (s,a,j) \in \mathcal{S}\times\mathcal{A}\times\{h\}; \\ 0, & \text{if } s = s_{\mathsf{aug}} \text{ or } j \neq h. \end{cases} \tag{35}$$

In addition, the augmented probability transition kernel $\widehat{P}^{\mathsf{aug},h}$ is constructed based on $\widehat{P}$ as follows:

$$\widehat{P}_j^{\mathsf{aug},h}(s' \,|\, s,a) = \begin{cases} \widehat{P}_j(s' \,|\, s,a), & \text{if } s' \in \mathcal{S} \\ 1 - \sum_{s'\in\mathcal{S}} \widehat{P}_j(s' \,|\, s,a), & \text{if } s' = s_{\mathsf{aug}} \end{cases} \qquad \text{for all } (s,a,j) \in \mathcal{S}\times\mathcal{A}\times[h]; \tag{36a}$$

$$\widehat{P}_j^{\mathsf{aug},h}(s' \,|\, s,a) = \mathbb{1}(s' = s_{\mathsf{aug}}) \qquad \text{if } s = s_{\mathsf{aug}} \text{ or } j > h. \tag{36b}$$

The other part of the procedure is nearly identical to the one for solving (20); see Algorithm 2 for details.

## 3.3 Intuition

Before proceeding to our main theory, we take a moment to explain the rationale behind our algorithm design. For simplicity of presentation, let us look at the special case where there exists a single fixed reward function of interest. Imagine that we are given an exploration policy

$$\pi_{\mathsf{b}} = \mathbb{E}_{\pi\sim\mu_{\mathsf{b}}}[\pi] \tag{37}$$

for some $\mu_{\mathsf{b}} \in \Delta(\Pi)$, which takes the form of some mixture of deterministic policies. We would like to sample $K$ episodes using this policy $\pi_{\mathsf{b}}$, and perform policy learning via an offline RL algorithm. In light of the state-of-the-art offline RL theory (Li et al., 2022b; Shi et al., 2022; Yin and Wang, 2021b), the total regret of a sample-efficient offline RL algorithm can be upper bounded (up to some logarithm factor) by

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \lesssim \sum_h \sum_{s,a} d_h^{\pi^{\star}}(s,a) \min\left\{ \sqrt{\frac{\mathsf{Var}_{P_{h,s,a}}\big(V_{h+1}^{\pi^{\star}}\big)}{K\mathbb{E}_{\pi'\sim\mu_{\mathsf{b}}}\big[d_h^{\pi'}(s,a)\big]}},\, H \right\}, \tag{38}$$

where $\pi^\star$ is the optimal policy, and $\widehat{\pi}$ represents the policy output by the offline RL algorithm. To further convert (38) into a more convenient upper bound, we make the observation that

$$\begin{aligned}
&\sum_h \sum_{s,a} d_h^{\pi^\star}(s,a) \min\left\{\sqrt{\frac{\mathsf{Var}_{P_{h,s,a}}\big(V_{h+1}^{\pi^\star}\big)}{K\mathbb{E}_{\pi'\sim\mu_{\mathsf{b}}}\big[d_h^{\pi'}(s,a)\big]}}, H\right\} \\
&\quad\le \sum_h \sum_{s,a} d_h^{\pi^\star}(s,a) \sqrt{\frac{\mathsf{Var}_{P_{h,s,a}}\big(V_{h+1}^{\pi^\star}\big)+H}{K\mathbb{E}_{\pi'\sim\mu_{\mathsf{b}}}\big[d_h^{\pi'}(s,a)\big]+1/H}} \\
&\quad\le \left[\sum_h \sum_{s,a} \frac{d_h^{\pi^\star}(s,a)}{1/H+K\mathbb{E}_{\pi'\sim\mu_{\mathsf{b}}}\big[d_h^{\pi'}(s,a)\big]} \cdot \sum_h \sum_{s,a} d_h^{\pi^\star}(s,a)\Big(\mathsf{Var}_{P_{h,s,a}}\big(V_{h+1}^{\pi^\star}\big)+H\Big)\right]^{\frac{1}{2}} \\
&\quad\le c_8 H \left[\sum_h \sum_{s,a} \frac{d_h^{\pi^\star}(s,a)}{1/H+K\mathbb{E}_{\pi'\sim\mu_{\mathsf{b}}}\big[d_h^{\pi'}(s,a)\big]}\right]^{\frac{1}{2}}
\end{aligned}$$

for some universal constant $c_8>0$, where the second line hold since $\min\left\{\frac{x}{y}+\frac{u}{w}\right\}\le\frac{x+u}{y+w}$ for any $x,y,u,w>0$, the penultimate line arises from Cauchy-Schwarz, and the last line is valid since (see Li et al. (2022b) or our analysis in Section 6)

$$\sum_h \sum_{s,a} d_h^{\pi^\star}(s,a)\mathsf{Var}_{P_{h,s,a}}\big(V_{h+1}^{\pi^\star}\big) \le O(H^2) \quad \text{and} \quad \sum_h \sum_{s,a} d_h^{\pi^\star}(s,a) H \le O(H^2).$$

Substitution into (38) then yields

$$\begin{aligned}
V^\star(\rho)-V^{\widehat{\pi}}(\rho) &\lesssim H\left[\sum_h \sum_{s,a} \frac{d_h^{\pi^\star}(s,a)}{1/H+K\mathbb{E}_{\pi'\sim\mu_{\mathsf{b}}}\big[\widehat{d}_h^{\pi'}(s,a)\big]}\right]^{\frac{1}{2}} \\
&\lesssim H\max_{\pi\in\Pi}\left[\sum_h \sum_{s,a} \frac{d_h^{\pi}(s,a)}{1/H+K\mathbb{E}_{\pi'\sim\mu_{\mathsf{b}}}\big[\widehat{d}_h^{\pi'}(s,a)\big]}\right]^{\frac{1}{2}}. \qquad (39)
\end{aligned}$$

Everything then comes down to optimizing the following quantity

$$\max_{\pi\in\Pi} \sum_h \sum_{s,a} \frac{d_h^{\pi}(s,a)}{\frac{1}{KH}+\mathbb{E}_{\pi'\sim\mu_{\mathsf{b}}}\big[\widehat{d}_h^{\pi'}(s,a)\big]}, \qquad (40)$$

provided that we have sufficiently accurate information about the occupancy distribution $d_h^\pi$ for every $\pi$.

As it turns out, by choosing $\mu_{\mathsf{b}}\in\Delta(\Pi)$ to be a solution to the following convex program:

$$\arg\max_{\mu\in\Delta(\Pi)} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \log\left[\frac{1}{KH}+\mathbb{E}_{\pi\sim\mu}\big[\widehat{d}_h^{\pi}(s,a)\big]\right], \qquad (41)$$

we can control the quantity (40) to be at the desired level, as implied by the following elementary fact.

**Lemma 2** (Kiefer and Wolfowitz (1960)). *By taking $\mu_{\mathsf{b}}$ to be a solution to (41), we have*

$$\max_{\pi\in\Pi} \sum_h \sum_{s,a} \frac{\widehat{d}_h^{\pi}(s,a)}{\frac{1}{KH}+\mathbb{E}_{\pi'\sim\mu_{\mathsf{b}}}\big[\widehat{d}_h^{\pi'}(s,a)\big]} \le HSA. \qquad (42)$$

This lemma is a direct consequence of the main theorem in Kiefer and Wolfowitz (1960). Combining Lemma 2 with (39), we arrive at

$$V^\star(\rho)-V^{\widehat{\pi}}(\rho) \le O\left(\sqrt{\frac{H^3SA}{K}}\right), \qquad (43)$$

thus attaining the desired order. This explains the rationale behind the subproblem (20), provided that $d_h^\pi$ can be estimated faithfully. The other subproblem (14) can be elucidated analogously, which we omit here.

# 4 Main results

We are now positioned to present our main theoretical guarantees for the proposed algorithm, and we begin by looking at the reward-agnostic scenario with at most a polynomial number of fixed reward functions.

**Theorem 1** (Reward-agnostic RL). *Consider any given $0 < \delta < 1$ and $0 < \varepsilon < 1$. Suppose that there are $m_{\mathsf{reward}} = \mathsf{poly}(H, S, A)$ fixed reward functions of interest. Using the same batch of collected data that obey*

$$K \geq c_K \frac{H^3 SA}{\varepsilon^2} \log \frac{KH}{\delta} \qquad \text{and} \qquad KH \geq N \geq c_N \sqrt{H^9 S^7 A^7 K} \log \frac{HSA}{\delta} \tag{44}$$

*for some sufficiently large universal constants $c_K, c_N > 0$, we can guarantee that with probability at least $1 - \delta$, the proposed algorithm (cf. Algorithm 1) is able to achieve*

$$V^\star(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon, \tag{45}$$

*for each of these reward functions.*

The proof of this theorem is deferred to Section 6. Remarkably, Theorem 1 establishes the sample complexity of the proposed algorithm for the reward-agnostic setting. More precisely, taking $N = c_N \sqrt{H^9 S^7 A^7 K} \log \frac{HSA}{\delta}$ in Theorem 1 reveals that: to yield an $\varepsilon$-optimal policy, it suffices for our algorithm to have a sample size as small as

$$\text{(sample complexity)} \qquad N_{\mathsf{tot}} = K + NH \leq 2K = \widetilde{O}\left(\frac{H^3 SA}{\varepsilon^2}\right) \text{ episodes}, \tag{46}$$

with the proviso that $K$ is sufficiently large (or equivalently, $\varepsilon$ is sufficiently small). Here, we recall that $N_{\mathsf{tot}} = K + NH$ is the total number of collected sample episodes. Encouragingly, this sample complexity is provably minimax-optimal up to logarithmic factor; to justify this, even in the reward-aware case with a single reward function of interest, it has been shown by Domingues et al. (2021); Jin et al. (2018) that the sample complexity cannot go below $\frac{H^3 SA}{\varepsilon^2}$ (up to logarithmic factor) regardless of the algorithm in use. To the best of our knowledge, the present paper offers the first algorithm that provably achieves minimax optimality in this scenario.

What is more, the exploration algorithm we develop turns out to be sample-efficient for the reward-free setting as well, the scenario where one would like to simultaneously account for arbitrary reward functions (including those designed adversarially). Our theoretical guarantees are as follows.

**Theorem 2** (Reward-free RL). *Consider any given $0 < \delta < 1$ and $0 < \varepsilon < 1$. Using the same batch of collected data that obey*

$$K \geq c_K \frac{H^3 S^2 A}{\varepsilon^2} \log \frac{KH}{\delta} \qquad \text{and} \qquad KH \geq N \geq c_N \sqrt{H^9 S^7 A^7 K} \log \frac{HSA}{\delta} \tag{47}$$

*for some sufficiently large universal constants $c_K, c_N > 0$, we can guarantee that with probability at least $1 - \delta$, the proposed algorithm (cf. Algorithm 1) is able to achieve*

$$V^\star(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon \tag{48}$$

*uniformly over all possible reward functions.*

The proof of this theorem is postponed to Section 7. In a nutshell, the proof is based on the analysis for the reward-agnostic case in conjuction with standard uniform concentration bounds. Theorem 2 asserts that the sample complexity required for the proposed algorithm to achieve $\varepsilon$-optimal policy is

$$\text{(sample complexity)} \qquad N_{\mathsf{tot}} = K + NH \leq 2K = \widetilde{O}\left(\frac{H^3 S^2 A}{\varepsilon^2}\right) \text{ episodes}, \tag{49}$$

provided that we take $N = c_N \sqrt{H^9 S^7 A^7 K} \log \frac{HSA}{\delta}$ and that $K$ is sufficiently large. In comparison to (46), this sample complexity (49) is $S$ times larger, due to the more stringent requirement to accommodate all

reward functions uniformly. Interestingly, this theory matches the state-of-the-art sample complexity result (i.e., Ménard et al. (2021a)) derived so far for this reward-free setting. Regard lower bounds, it has been previously shown in Jin et al. (2020a) that a sample size on the order of $H^2S^2A/\varepsilon^2$ (up to logarithmic factor) is necessary for *stationary* finite-horizon MDPs; it is widely conjectured that the minimax lower bound for the non-stationary case studied herein should be a factor of $H$ larger than the lower limit for the stationary counterpart.

# 5 Prior art

In this section, we briefly overview a subset of other related works.

**Reward-aware exploration.** The studies of online exploration have been a central topic in RL. Here, we mention in passing a small number of representative works. The development of the UCRL algorithm, the UCRL2 algorithm and their variants (Auer and Ortner, 2006; Jaksch et al., 2010) exemplified earlier effort in implementing the optimism principle in the face of uncertainty. A more sample-efficient model-based online RL algorithm, called *upper confidence bound value iteration (UCBVI)*, was later proposed by Azar et al. (2017), which yields minimax-optimal regret asymptotically. Turning to model-free algorithms, Jin et al. (2018) justified the efficacy of Q-learning (in conjunction with UCB-type exploration) in online RL, which achieves a regret that is a factor of $\sqrt{H}$ away from optimal if the Bernstein-type confidence bounds are employed. Other versions of Q-learning-type algorithms, including the ones that come with low switching cost and the ones tailored to discounted infinite-horizon MDPs, have since been developed (Bai et al., 2019; Dong et al., 2019), among which a variance-reduced variant achieves asymptotically optimal regret (Zhang et al., 2020b). Two recent works Li et al. (2022a); Ménard et al. (2021b) further demonstrated how to reduce the burn-in cost — the sample size required to attain sample optimality — to be linear (and hence minimal) in $SA$. All these algorithms, however, rely on *a priori* reward information, which are ill-suited to the reward-agnostic convext. Additionally, information-theoretic regret lower bounds for reward-aware online RL were first developed by Jaksch et al. (2010); Jin et al. (2018), and revisited later on by Domingues et al. (2021). Going beyond the tabular case, a number of papers have further pursued efficient reward-aware exploration in the presence of low-dimensional function approximation (e.g., Ayoub et al. (2020); Du et al. (2021); Jin et al. (2020b); Li et al. (2021)).

**Reward-free and task-agnostic exploration.** Moving on to reward-free exploration (the case where one is asked to account for an arbitrarily large number of reward functions), the R-max type algorithm proposed in the earlier work Brafman and Tennenholtz (2002) incurs a sample complexity at least as large as $\frac{H^{11}S^2A}{\varepsilon^3}$ (see Jin et al. (2020a, Appendix A)). Recently, Jin et al. (2020a) came up with a clever scheme — with instantaneous rewards chosen to be indicator functions regarding state/action visitations — that is guaranteed to work with $\widetilde{O}\big(\frac{H^5S^2A}{\varepsilon^2}\big)$ sample episodes. This sample complexity bound is further improved by Kaufmann et al. (2021); Ménard et al. (2021a), with Ménard et al. (2021a) designing an algorithm with optimal sample complexity (i.e., $\widetilde{O}\big(\frac{H^3S^2A}{\varepsilon^2}\big)$) in the reward-free setting. Another work Zhang et al. (2021b) studied a different setting with "totally bounded rewards" (so that the sum of immediate rewards over all steps is bounded above by 1) in stationary finite-horizon MDPs; when translated to the bounded reward setting (so that any immediate reward is bounded above by 1), the algorithm put forward in Zhang et al. (2021b) exhibited a sample size of $\widetilde{O}\big(\frac{H^2S^2A}{\varepsilon^2}\big)$ episodes.[3] Reward-free exploration with low policy switching cost has been investigated in Qiao et al. (2022), while the effect of low-dimensional model representation in RFE has been further explored in Chen et al. (2022, 2021); Qiao and Wang (2022); Qiu et al. (2021); Wagenmaker et al. (2022); Wang et al. (2020); Zanette et al. (2020); Zhang et al. (2021a). It is also related to the problem of uniform policy evaluation, which aims to ensure reliable policy evaluation uniformly over all policies (Yin and Wang, 2021a). In contrast to the reward-free setting that covers arbitrarily many reward functions, Zhang et al. (2020a) assumed the existence of only $N$ reward functions of interest, and proposed a model-free algorithm that enjoys a sample complexity of $\widetilde{O}\big(\frac{H^5SA\log m_{\mathsf{reward}}}{\varepsilon^2}\big)$. This result, however, is suboptimal in terms of the horizon dependency. Finally, reward-free exploration has also been studied in

[3]Note that this horizon dependency $H^2$ (proven for homogeneous MDPs) is in general not possible in non-stationary MDPs.

the context of safe RL (Huang et al., 2022) and constrained RL (Miryoosefi and Jin, 2022), which are beyond the scope of the current paper.

**Offline RL.** The past few years have seen much activity in offline RL (also called batch RL) (Levine et al., 2020). The principle of pessimism (or conservatism) in the face of uncertainty has been shown to be effective in solving offline RL (Jin et al., 2022, 2021; Kumar et al., 2020; Li et al., 2022b; Rashidinejad et al., 2021). The sample complexity of offline RL in the tabular case has been tackled by a recent line of works (Li et al., 2022b; Rashidinejad et al., 2021; Shi et al., 2022; Xie et al., 2021; Yan et al., 2022a; Yin et al., 2021; Yin and Wang, 2021b); take finite-horizon MDPs with non-stationary transition kernels for example: the minimax-optimal sample complexity for attaining $\varepsilon$-accuracy is shown to be $\frac{H^3SC^\star}{\varepsilon^2}$ episodes for any $\varepsilon \in (0, H]$, where $C^\star$ stands for some single-policy concentrability coefficient that captures the resulting distribution shift between the optimal policy and the behavior policy (Li et al., 2022b). In particular, sample optimality for the full $\varepsilon$-range is achievable via the model-based approach (Li et al., 2022b), and establishing this result relies on modern statistical tools like the leave-one-out decoupling argument (Agarwal et al., 2020; Li et al., 2023). Sample-efficient offline RL algorithms have also been studied in more complicated scenarios, including but not limited to the case with linear function approximation (Jin et al., 2021; Xu and Liang, 2022) and zero-sum Markov games (Cui and Du, 2022a,b; Yan et al., 2022b).

# 6 Analysis for reward-agnostic exploration (proof of Theorem 1)

In this section, we present the proof of our main result for reward-agnostic exploration in Theorem 1. Towards this end, we shall first establish the following result when there is a single reward function of interest (i.e., $m_{\mathsf{reward}} = 1$).

**Theorem 3.** *Consider any given $0 < \delta < 1$ and $0 < \varepsilon < 1$. Suppose that there is only $m_{\mathsf{reward}} = 1$ reward function of interest and it is independent of the data samples. With probability at least $1 - \delta$, the proposed algorithm (cf. Algorithm 1) achieves*

$$V^\star(\rho) - V^{\widehat{\pi}}(\rho) \le \varepsilon, \tag{50}$$

*provided that*

$$K \ge c_K \frac{H^3SA}{\varepsilon^2} \log \frac{KH}{\delta} \qquad \text{and} \qquad KH \ge N \ge c_N \sqrt{H^9S^7A^7K} \log \frac{HSA}{\delta} \tag{51}$$

*for some sufficiently large universal constants $c_K, c_N > 0$.*

As it turns out, Theorem 1 (with $m_{\mathsf{reward}} = \mathsf{poly}(H, S, A)$) is a direct consequence of Theorem 1. To see this, replacing $\delta$ with $\delta / m_{\mathsf{reward}} \asymp \delta / \mathsf{poly}(H, S, A)$ in Theorem 3 and taking the union bound over all $m_{\mathsf{reward}}$ reward functions of interest suffice to justify Theorem 1. Consequently, the remainder of this section is dedicated to proving Theorem 3.

Before continuing, we isolate several additional notation that might be useful in presenting our proof. We introduce the following vector notation for value functions: for any $1 \le h \le H$,

$$V_h^\pi = \left[V_h^\pi(s)\right]_{s\in\mathcal{S}} \quad \text{for any policy } \pi \qquad \text{and} \qquad V_h^\star = \left[V_h^\star(s)\right]_{s\in\mathcal{S}}. \tag{52}$$

We shall also use $\widehat{V}_h = [\widehat{V}_h(s)]_{s\in\mathcal{S}}$ to represent the value estimate for step $h$ by Algorithm 4. The state occupancy distribution for policy $\pi$ at step $h \in [H]$ is also represented by the following vector

$$d_h^\pi = \left[d_h^\pi(s)\right]_{s\in\mathcal{S}}. \tag{53}$$

We also use the shorthand notation $P_{h,s,a} \in \mathbb{R}^{1\times S}$ (resp. $\widehat{P}_{h,s,a} \in \mathbb{R}^{1\times S}$) to denote $P_h(\cdot \mid s, a)$ (resp. $\widehat{P}_h(\cdot \mid s, a)$).

## 6.1 Preliminary facts

Before embarking on the proofs of our main theory, we first gather several useful preliminary results.

**A direct consequence of the stopping rule.** To begin with, recall the stopping rule of Algorithm 3 (see line 6) and the optimality of $\pi^{(t)}$ (see line 3 in Algorithm 3 and also the discussion at the end of Section 3.2). These taken collectively allow one to demonstrate that for any policy $\pi$,

$$\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\widehat{d}_h^{\pi}(s,a)}{\frac{1}{KH}+\mathbb{E}_{\pi'\sim\widehat{\mu}_{\mathsf{b}}}\big[\widehat{d}_h^{\pi'}(s,a)\big]}=\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\widehat{d}_h^{\pi^{(t)}}(s,a)}{\frac{1}{KH}+\mathbb{E}_{\pi'\sim\widehat{\mu}_{\mathsf{b}}}\big[\widehat{d}_h^{\pi'}(s,a)\big]}$$
$$=g(\pi^{(t)},\widehat{d},\widehat{\mu}_{\mathsf{b}})\le 2HSA, \tag{54}$$

where $\widehat{\mu}_{\mathsf{b}}$ is the output returned by Algorithm 3, and $\pi^{(t)}$ represents the optimal policy computed right before the termination of this algorithm. Here, the first relation arises since the left-hand side is the value function of the MDP $\mathcal{M}_{\mathsf{b}}$ w.r.t. policy $\pi$, which is maximized by $\pi^{(t)}$; the last inequality is due to the stopping rule. This property (54), which results from our carefully designed exploration scheme, plays a key role in the subsequent analysis.

**Properties about the model-based offline algorithm.** Next, we collect a couple of preliminary facts that have been previously established in Li et al. (2022b) for the model-based algorithm in Algorithm 4.

**Lemma 3.** *With probability exceeding $1-\delta$, the policy $\widehat{\pi}$ returned by Algorithm 4 obeys*

$$\big\langle d_h^{\pi^\star},V_h^\star-V_h^{\widehat{\pi}}\big\rangle\le\big\langle d_h^{\pi^\star},V_h^\star-\widehat{V}_h\big\rangle\le 2\sum_{j:j\ge h}\sum_{s\in\mathcal{S}}d_j^{\pi^\star}(s)b_j\big(s,\pi_j^\star(s)\big)\le 2\sum_{j:j\ge h}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_j^{\pi^\star}(s,a)b_j(s,a) \tag{55}$$

*for all $1\le h\le H$. Additionally, with probability exceeding $1-\delta$, one has*

$$\mathsf{Var}_{\widehat{P}_{h,s,a}}(\widehat{V}_{h+1})\le 2\mathsf{Var}_{P_{h,s,a}}(\widehat{V}_{h+1})+\frac{5H^2\log\frac{KH}{\delta}}{\widehat{N}_h^{\mathsf{b}}(s,a)},\qquad\forall(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]. \tag{56}$$

*Here, $\widehat{P}=\{\widehat{P}_h\}_{1\le h\le H}$ denotes the empirical transition kernel constructed in Algorithm 4.*

*Proof.* The claim (55) follows immediately from Li et al. (2022b, Eqn. (126)). The claim (56) follows from Li et al. (2022b, Lemma 8) and the fact that $\widehat{V}_{h+1}$ is independent of $\widehat{P}_{h,s,a}$ (owing to the algorithm design; see Li et al. (2022b, Section 5.2)). □

## 6.2 Main steps

**Step 1: accuracy of estimated occupancy distributions.** We first provide estimation guarantees for the estimated occupancy distributions $\widehat{d}_h^{\pi}$ (see (23) and (24) in Algorithm 1). The proof is deferred to Appendix 6.3.

**Lemma 4.** *Recall that $\xi=c_\xi H^3S^3A^3\log\frac{HSA}{\delta}$ for some large enough constant $c_\xi>0$ (see (18)). With probability at least $1-\delta$, the estimated occupancy distributions specified in (23) and (24) of Algorithm 1 satisfy*

$$\frac{1}{2}\widehat{d}_h^{\pi}(s,a)-\frac{\xi}{4N}\le d_h^{\pi}(s,a)\le 2\widehat{d}_h^{\pi}(s,a)+2e_h^{\pi}(s,a)+\frac{\xi}{4N} \tag{57}$$

*simultaneously for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$ and all deterministic Markov policy $\pi\in\Pi$, provided that*

$$KH\ge N\ge C_N\sqrt{H^9S^7A^7K}\log\frac{HSA}{\delta}\qquad\text{and}\qquad K\ge C_KHSA \tag{58}$$

*for some large enough constants $C_N,C_K>0$. Here, $\{e_h^{\pi}(s,a)\}$ is some non-negative sequence satisfying*

$$\sum_{s,a}e_h^{\pi}(s,a)\le\frac{2SA}{K}+\frac{13SAH\xi}{N}\lesssim\sqrt{\frac{SA}{HK}}\qquad\text{for all } h\in[H]\text{ and all deterministic Markov policy }\pi. \tag{59}$$

In a nutshell, this lemma makes apparent the effectiveness of our exploration stage: for all deterministic Markov policy $\pi$, we have obtained reasonably accurate estimation of the associated occupancy distributions. In particular, the estimation error for each state-action pair is inversely proportional to the number of episodes $N$ in each round, up to some additional error terms $\{e_h^\pi(s,a)\}$ whose aggregate contributions are well-controlled.

Before proceeding, we single out a direct consequence of Lemma 4 that will prove useful. Denoting by $N_h^{\mathsf{b}}(s,a)$ the number of visits to $(s,a)$ at step $h$ during Stage 1.2 of the proposed algorithm (which employs the behavior policy $\widehat{\pi}_{\mathsf{b}} = \mathbb{E}_{\pi\sim\widehat{\mu}_{\mathsf{b}}}[\pi]$ to sample $K$ episodes), we have

$$N_h^{\mathsf{b}}(s,a) \geq \widehat{N}_h^{\mathsf{b}}(s,a), \qquad \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H] \tag{60}$$

with probability exceeding $1-\delta$, where we remind the reader of the definition of $\widehat{N}_h^{\mathsf{b}}(s,a)$ in (22). This property is crucial when invoking Algorithm 4.

*Proof.* It is first seen from Lemma 4 that

$$\mathbb{E}[N_h^{\mathsf{b}}(s,a)] = K \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}_{\mathsf{b}}}\big[d_h^\pi(s,a)\big] \geq \frac{K}{2} \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}_{\mathsf{b}}}\big[\widehat{d}_h^\pi(s,a)\big] - \frac{K\xi}{4N}. \tag{61}$$

The Bernstein inequality together the union bound then implies that, with probability at least $1-\delta$,

$$\begin{aligned}
N_h^{\mathsf{b}}(s,a) &\geq \max\left\{\mathbb{E}\big[N_h^{\mathsf{b}}(s,a)\big] - \sqrt{4\mathbb{E}\big[N_h^{\mathsf{b}}(s,a)\big]\log\frac{HSA}{\delta}} - \log\frac{HSA}{\delta},\ 0\right\} \\
&\geq \max\left\{\mathbb{E}\big[N_h^{\mathsf{b}}(s,a)\big] - \frac{1}{2}\mathbb{E}\big[N_h^{\mathsf{b}}(s,a)\big] - 2\log\frac{HSA}{\delta} - \log\frac{HSA}{\delta},\ 0\right\} \\
&\geq \max\left\{\frac{K}{4}\mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}_{\mathsf{b}}}\big[\widehat{d}_h^\pi(s,a)\big] - \frac{K\xi}{8N} - 3\log\frac{HSA}{\delta},\ 0\right\} = \widehat{N}_h^{\mathsf{b}}(s,a)
\end{aligned}$$

holds simultaneously over all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, where the second line invokes the AM-GM inequality, and the last line makes use of (61). □

**Step 2: bounding $V_j^\star - V_j^{\widehat{\pi}}$ for step $j$.** We now turn to the sub-optimality of the final policy estimate $\widehat{\pi}$. To begin with, we define, for each $1\leq h\leq H$, the following set

$$\mathcal{I}_h := \left\{(s,a) : \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}_{\mathsf{b}}}\big[\widehat{d}_h^\pi(s,a)\big] \geq \frac{4\xi}{N}\right\}, \tag{62}$$

containing all state-action pairs whose associated empirical occupancy density (weighted over the sampling policy $\widehat{\mu}_{\mathsf{b}}$) exceed some prescribed threshold. For each $1\leq j\leq H$, one can then bound and decompose

$$\begin{aligned}
\big\langle d_j^{\pi^\star}, V_j^\star - V_j^{\widehat{\pi}}\big\rangle \leq \big\langle d_j^{\pi^\star}, V_j^\star - \widehat{V}_j\big\rangle &\overset{\text{(i)}}{\leq} 2\sum_{h:h\geq j}\sum_{s,a} d_h^{\pi^\star}(s,a) b_h(s,a) \\
&\overset{\text{(ii)}}{\leq} 2\sum_{h:h\geq j}\sum_{s,a}\left(2\widehat{d}_h^{\pi^\star}(s,a) + 2e_{h+1}^{\pi^\star}(s,a) + \frac{\xi}{N}\right) b_h(s,a) \\
&\overset{\text{(iii)}}{\lesssim} \sum_{h:h\geq j}\sum_{s,a}\widehat{d}_h^{\pi^\star}(s,a) b_h(s,a) + H^2\left(\sqrt{\frac{SA}{HK}} + \frac{SA\xi}{N}\right) \\
&\leq \underbrace{\sum_{h:h\geq j}\sum_{(s,a)\in\mathcal{I}_h}\widehat{d}_h^{\pi^\star}(s,a) b_h(s,a)}_{=:\alpha_1} + \underbrace{\sum_{h:h\geq j}\sum_{(s,a)\notin\mathcal{I}_h}\widehat{d}_h^{\pi^\star}(s,a) b_h(s,a)}_{=:\alpha_2} + H^2\left(\sqrt{\frac{SA}{HK}} + \frac{SA\xi}{N}\right).
\end{aligned} \tag{63}$$

Here, (i) is a consequence of property (55) in Lemma 3, (ii) results from (57) in Lemma 4 that quantifies the discrepancy between $d_h^{\pi^\star}$ and $\widehat{d}_h^{\pi^\star}$, whereas (iii) follows from (59) in Lemma 4 as well as the basic fact that $b_h(s,a)\leq H$ (see (105)). We shall then cope with the two terms $\alpha_1$ and $\alpha_2$ separately.

- With regards to the first term $\alpha_1$ in (63), it follows from the choice (105) of $b_h$ that

$$\begin{aligned}
\alpha_1 &\lesssim \sum_{h:h\geq j}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi^\star}(s,a)\min\left\{\sqrt{\frac{\log\frac{KH}{\delta}}{\widehat{N}_h^{\mathsf{b}}(s,a)}\mathsf{Var}_{\widehat{P}_{h,s,a}}\big(\widehat{V}_{h+1}\big)}+H\frac{\log\frac{KH}{\delta}}{\widehat{N}_h^{\mathsf{b}}(s,a)},\ H\right\}\\
&\overset{\text{(i)}}{\lesssim} \sum_{h:h\geq j}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi^\star}(s,a)\min\left\{\sqrt{\frac{\log\frac{KH}{\delta}}{\widehat{N}_h^{\mathsf{b}}(s,a)}\mathsf{Var}_{P_{h,s,a}}\big(\widehat{V}_{h+1}\big)}+H\frac{\log\frac{KH}{\delta}}{\widehat{N}_h^{\mathsf{b}}(s,a)},\ H\right\}\\
&\leq \sum_{h:h\geq j}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi^\star}(s,a)\left\{\min\left\{\sqrt{\frac{\mathsf{Var}_{P_{h,s,a}}\big(\widehat{V}_{h+1}\big)\log\frac{KH}{\delta}}{\widehat{N}_h^{\mathsf{b}}(s,a)}},\ H\right\}+H\frac{\log\frac{KH}{\delta}}{\widehat{N}_h^{\mathsf{b}}(s,a)}\right\}\\
&\overset{\text{(ii)}}{\lesssim} \sum_{h:h\geq j}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi^\star}(s,a)\left\{\sqrt{\frac{\mathsf{Var}_{P_{h,s,a}}\big(\widehat{V}_{h+1}\big)\log\frac{KH}{\delta}+H}{\widehat{N}_h^{\mathsf{b}}(s,a)+1/H}}+H\frac{\log\frac{KH}{\delta}}{\widehat{N}_h^{\mathsf{b}}(s,a)}\right\}\\
&\overset{\text{(iii)}}{\lesssim} \sum_{h:h\geq j}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi^\star}(s,a)\left\{\sqrt{\frac{\mathsf{Var}_{P_{h,s,a}}\big(\widehat{V}_{h+1}\big)\log\frac{KH}{\delta}+H}{K\mathbb{E}_{\pi\sim\widehat{\mu}_{\mathsf{b}}}\big[\widehat{d}_h^{\pi}(s,a)\big]+1/H}}+\frac{H\log\frac{KH}{\delta}}{1/H+K\mathbb{E}_{\pi\sim\widehat{\mu}_{\mathsf{b}}}\big[\widehat{d}_h^{\pi}(s,a)\big]}\right\}\\
&\overset{\text{(iv)}}{\lesssim} \left\{\sum_{h:h\geq j}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi^\star}(s,a)\sqrt{\frac{\mathsf{Var}_{P_{h,s,a}}\big(\widehat{V}_{h+1}\big)\log\frac{KH}{\delta}+H}{1/H+K\mathbb{E}_{\pi\sim\widehat{\mu}_{\mathsf{b}}}\big[\widehat{d}_h^{\pi}(s,a)\big]}}\right\}+\frac{H^2SA\log\frac{KH}{\delta}}{K}.
\end{aligned}\tag{64}$$

  Here, (i) makes use of property (56) in Lemma 3, (ii) is valid due to the elementary inequality $\min\{\frac{x}{y},\frac{u}{w}\}\leq\frac{x+u}{y+w}$ for any $x,y,u,w>0$, (iii) follows since

$$\min_{(s,a)\in\mathcal{I}_h}\widehat{N}_h^{\mathsf{b}}(s,a)\gtrsim K\mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}_{\mathsf{b}}}\big[\widehat{d}_h^{\pi}(s,a)\big]+\frac{1}{H}$$

  holds according to (22) and (62), while (iv) results from property (54).

- When it comes to the other term $\alpha_2$ in (63), we observe that

$$\begin{aligned}
\sum_{h:h\geq j}\sum_{(s,a)\notin\mathcal{I}_h}\widehat{d}_h^{\pi^\star}(s,a) &\leq \sum_{h:h\geq j}\sum_{(s,a)\notin\mathcal{I}_h}\frac{\big(\frac{1}{KH}+\frac{4\xi}{N}\big)\widehat{d}_h^{\pi^\star}(s,a)}{\frac{1}{KH}+\mathbb{E}_{\pi\sim\widehat{\mu}_{\mathsf{b}}}\big[\widehat{d}_h^{\pi}(s,a)\big]}\\
&\leq\left(\frac{1}{KH}+\frac{4\xi}{N}\right)2HSA\\
&\asymp\frac{HSA\xi}{N},
\end{aligned}$$

  where the first line follows from the definition (62) of $\mathcal{I}_h$, the second line relies on (54), and the last line is valid since $\frac{1}{KH}\lesssim\frac{\xi}{N}$ holds as long as $N\leq KH$ and $\xi\geq 1$. This taken together with the fact that $b_h(s,a)\leq H$ (see (105)) immediately gives

$$\alpha_2\lesssim\frac{H^2SA\xi}{N}.\tag{65}$$

Substituting the preceding bounds (64) and (65) on $\alpha_1$ and $\alpha_2$ into (63), we arrive at

$$\begin{aligned}
\big\langle d_j^{\pi^\star},V_j^\star-V_j^{\widehat{\pi}}\big\rangle &\leq \big\langle d_j^{\pi^\star},V_j^\star-\widehat{V}_j\big\rangle\lesssim\alpha_1+\alpha_2+H^2\Big(\sqrt{\frac{SA}{HK}}+\frac{SA\xi}{N}\Big)\\
&\lesssim\underbrace{\sum_{h:h\geq j}\sum_{(s,a)\in\mathcal{I}_h}\widehat{d}_h^{\pi^\star}(s,a)\sqrt{\frac{\mathsf{Var}_{P_{h,s,a}}\big(\widehat{V}_{h+1}\big)\log\frac{KH}{\delta}+H}{1/H+K\mathbb{E}_{\pi\sim\widehat{\mu}_{\mathsf{b}}}\big[\widehat{d}_h^{\pi}(s,a)\big]}}}_{=:\beta}+\frac{H^2SA\log\frac{KH}{\delta}}{K}+H^2\Big(\sqrt{\frac{SA}{HK}}+\frac{SA\xi}{N}\Big).
\end{aligned}\tag{66}$$

Everything then comes down to controlling the term $\beta$, which forms the main content of the remaining proof.

**Step 3: controlling the term $\beta$.** To bound the term $\beta$ in (66), we first apply the Cauchy-Schwarz inequality to reach

$$\beta \le \underbrace{\left(\sum_{h:h\ge j}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi^\star}(s,a)\left[\mathsf{Var}_{P_{h,s,a}}\big(\widehat{V}_{h+1}\big)\log\frac{KH}{\delta}+H\right]\right)^{1/2}}_{=:\beta_1}\underbrace{\left(\sum_{h:h\ge j}\sum_{(s,a)\in\mathcal{I}_h}\frac{\widehat{d}_h^{\pi^\star}(s,a)}{1/H+K\mathbb{E}_{\pi\sim\widehat{\mu}_{\mathsf{b}}}\big[\widehat{d}_h^{\pi}(s,a)\big]}\right)^{1/2}}_{=:\beta_2}, \tag{67}$$

thus motivating us to bound $\beta_1$ and $\beta_2$ separately.

- Regarding the first term $\beta_1$, we make the observation that

$$\begin{aligned}
(\beta_1)^2 &\le \sum_h\sum_{(s,a)}\widehat{d}_h^{\pi^\star}(s,a)\left\{\mathsf{Var}_{P_{h,s,a}}\big(\widehat{V}_{h+1}\big)\log\frac{KH}{\delta}\right\}+\sum_h\sum_{(s,a)}\widehat{d}_h^{\pi^\star}(s,a)H\\
&\le \sum_h\sum_{(s,a)}\left(2d_h^{\pi^\star}(s,a)+\frac{2\xi}{N}\right)\mathsf{Var}_{P_{h,s,a}}\big(\widehat{V}_{h+1}\big)\log\frac{KH}{\delta}+H^2\\
&\le 2\sum_h\sum_{(s,a)}d_h^{\pi^\star}(s,a)\Big\{\mathsf{Var}_{P_{h,s,a}}\big(V_{h+1}^\star\big)+\mathsf{Var}_{P_{h,s,a}}\big(V_{h+1}^\star-\widehat{V}_{h+1}\big)\Big\}\log\frac{KH}{\delta}+\frac{2\xi}{N}SAH^3\log\frac{KH}{\delta}+H^2\\
&\lesssim \underbrace{\sum_h\sum_{(s,a)}d_h^{\pi^\star}(s,a)\mathsf{Var}_{P_{h,s,a}}\big(V_{h+1}^\star\big)}_{=:\beta_{1,1}}\log\frac{KH}{\delta}+\underbrace{\sum_h\sum_{(s,a)}d_h^{\pi^\star}(s,a)\mathsf{Var}_{P_{h,s,a}}\big(V_{h+1}^\star-\widehat{V}_{h+1}\big)}_{=:\beta_{1,2}}\log\frac{KH}{\delta}\\
&\quad+\frac{\xi}{N}SAH^3\log\frac{KH}{\delta}+H^2,
\end{aligned}$$

  where the second line invokes Lemma 4 and the fact $\sum_{(s,a)}\widehat{d}_h^{\pi^\star}(s,a)\le 1$, and the third line is valid since $\mathsf{Var}(X+Y)\le 2\mathsf{Var}(X)+2\mathsf{Var}(Y)$.

  - To cope with the first term $\beta_{1,1}$ in the above display, we find it helpful to define, for each $1\le h\le H$, a distribution vector $d_h^{\pi^\star}=\big[d_h^{\pi^\star}(s)\big]_{s\in\mathcal{S}}\in\mathbb{R}^S$, a reward vector $r_h^{\pi^\star}=\big[r_h\big(s,\pi_h^\star(s)\big)\big]_{s\in\mathcal{S}}\in\mathbb{R}^S$, and a matrix $P_h^\star\in\mathbb{R}^{S\times S}$ obeying

$$P_h^\star(s,s')=P_h\big(s'\,|\,s,\pi_h^\star(s)\big),\qquad s,s'\in\mathcal{S}.$$

    Given that $\pi^\star$ is a deterministic policy, one can write

$$\begin{aligned}
\sum_{(s,a)}d_h^{\pi^\star}(s,a)\mathsf{Var}_{P_{h,s,a}}\big(V_{h+1}^\star\big)&=\sum_s d_h^{\pi^\star}\big(s,\pi_h^\star(s)\big)\mathsf{Var}_{P_{h,s,\pi_h^\star(s)}}\big(V_{h+1}^\star\big)\\
&=\big\langle d_h^{\pi^\star},P_h^\star(V_{h+1}^\star\circ V_{h+1}^\star)-(P_h^\star V_{h+1}^\star)\circ(P_h^\star V_{h+1}^\star)\big\rangle.
\end{aligned}$$

    This allows one to deduce that

$$\begin{aligned}
\beta_{1,1}&=\sum_{h=1}^H\big\langle d_h^{\pi^\star},P_h^\star(V_{h+1}^\star\circ V_{h+1}^\star)-(P_h^\star V_{h+1}^\star)\circ(P_h^\star V_{h+1}^\star)\big\rangle\\
&\overset{\text{(i)}}{=}\sum_{h=1}^H\big\langle d_{h+1}^{\pi^\star},V_{h+1}^\star\circ V_{h+1}^\star\big\rangle-\sum_{h=1}^H\big\langle d_h^{\pi^\star},(P_h^\star V_{h+1}^\star)\circ(P_h^\star V_{h+1}^\star)\big\rangle\\
&\overset{\text{(ii)}}{=}\sum_{h=1}^H\big\langle d_{h+1}^{\pi^\star},V_{h+1}^\star\circ V_{h+1}^\star\big\rangle-\sum_{h=1}^H\big\langle d_h^{\pi^\star},V_h^\star\circ V_h^\star\big\rangle+2\sum_{h=1}^H\big\langle d_h^{\pi^\star},r_h^{\pi^\star}\circ(P_h^\star V_{h+1}^\star)\big\rangle-\sum_{h=1}^H\big\langle d_h^{\pi^\star},r_h^\star\circ r_h^\star\big\rangle\\
&\le\sum_{h=1}^H\big\langle d_{h+1}^{\pi^\star},V_{h+1}^\star\circ V_{h+1}^\star\big\rangle-\sum_{h=1}^H\big\langle d_h^{\pi^\star},V_h^\star\circ V_h^\star\big\rangle+2\sum_{h=1}^H\big\langle d_h^{\pi^\star},r_h^{\pi^\star}\circ(P_h^\star V_{h+1}^\star)\big\rangle
\end{aligned}$$

$$\overset{\text{(iii)}}{\leq} 2\sum_{h=1}^{H}\big\langle d_h^{\pi^\star}, r_h^{\pi^\star}\circ(P_h^\star V_{h+1}^\star)\big\rangle \overset{\text{(iv)}}{\leq} 2H^2, \tag{68}$$

where (i) follows from $(d_h^{\pi^\star})^\top P_h^\star = (d_{h+1}^{\pi^\star})^\top$, (ii) makes use of the Bellman equation $V_h^\star = r_h + P_h^\star V_{h+1}^\star$, (iii) results from the telescoping sum and the fact that $V_{H+1}^\star = 0$, and (iv) holds since $V_h^\star(s) \leq H$ and $r_h(s,a) \leq 1$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.

– Regarding $\beta_{1,2}$, we first derive a useful crude bound as follows:

$$\begin{aligned}
\big\langle d_j^{\pi^\star}, V_j^\star - \widehat{V}_j\big\rangle &\lesssim \sum_{h:h\geq j}\sum_{(s,a)\in\mathcal{I}_h}\frac{H\widehat{d}_h^{\pi^\star}(s,a)\sqrt{\log\frac{KH}{\delta}}}{\sqrt{1/H + K\mathbb{E}_{\pi\sim\widehat{\mu}_{\mathrm{b}}}\big[\widehat{d}_h^{\pi}(s,a)\big]}} + \frac{H^2SA\log\frac{KH}{\delta}}{K} + H^2\Big(\sqrt{\frac{SA}{HK}} + \frac{SA\xi}{N}\Big)\\
&\lesssim H\sqrt{\log\frac{KH}{\delta}}\left\{\sum_h\sum_{(s,a)}\widehat{d}_h^{\pi^\star}(s,a)\right\}^{\frac{1}{2}}\left[\sum_h\sum_{(s,a)}\frac{\widehat{d}_h^{\pi^\star}(s,a)}{1/H + K\mathbb{E}_{\pi'\sim\widehat{\mu}_{\mathrm{b}}}\big[\widehat{d}_h^{\pi'}(s,a)\big]}\right]^{\frac{1}{2}}\\
&\quad + \frac{H^2SA\log\frac{KH}{\delta}}{K} + H^2\Big(\sqrt{\frac{SA}{HK}} + \frac{SA\xi}{N}\Big)\\
&\lesssim \sqrt{\frac{H^4SA\log\frac{KH}{\delta}}{K}} + \frac{H^2SA\xi}{N} + \frac{H^2SA\log\frac{KH}{\delta}}{K} \lesssim 1.
\end{aligned} \tag{69}$$

Here, the first inequality follows from (66) and the fact that $\mathsf{Var}_{P_{h,s,a}}\big(\widehat{V}_{h+1}\big) \leq H^2$; the second line applies the Cauchy-Schwarz inequality; the third inequality arises from property (54) and the fact $\sum_{(s,a)}\widehat{d}_h^{\pi^\star}(s,a) \leq 1$; and the last inequality is valid as long as $K \gtrsim H^4SA\log\frac{KH}{\delta}$ and $N \gtrsim H^2SA\xi$. With this bound in mind, one can then control $\beta_{1,2}$ as follows

$$\begin{aligned}
\beta_{1,2} &\leq \sum_h\sum_{(s,a)} d_h^{\pi^\star}(s,a)\mathbb{E}_{P_{h,s,a}}\big[\big(V_{h+1}^\star - \widehat{V}_{h+1}\big)\circ\big(V_{h+1}^\star - \widehat{V}_{h+1}\big)\big]\\
&\leq H\sum_h\sum_{(s,a)} d_h^{\pi^\star}(s,a)\mathbb{E}_{P_{h,s,a}}\big[V_{h+1}^\star - \widehat{V}_{h+1}\big]\\
&= H\sum_h\big\langle d_{h+1}^{\pi^\star}, V_{h+1}^\star - \widehat{V}_{h+1}\big\rangle \leq H^2,
\end{aligned} \tag{70}$$

where the second line holds since $V_{h+1}^\star, \widehat{V}_{h+1} \in [0,H]$, the third line relies on the fact that $(d_h^{\pi^\star})^\top P_h^\star = (d_{h+1}^{\pi^\star})^\top$, and the last relation follows from (69).

Therefore, the above bounds (68) and (70) allow one to readily conclude that

$$(\beta_1)^2 \lesssim (\beta_{1,1}+\beta_{1,2})\log\frac{KH}{\delta} + \frac{\xi}{N}SAH^3 + H^2 \lesssim H^2\log\frac{KH}{\delta}, \tag{71}$$

with the proviso that $N \gtrsim \xi SAH$.

- When it comes to $\beta_2$, we can invoke (54) to yield

$$(\beta_2)^2 \leq \sum_{h:h\geq j}\sum_{(s,a)\in\mathcal{I}_h}\frac{\widehat{d}_h^{\pi^\star}(s,a)}{1/H + K\mathbb{E}_{\pi\sim\widehat{\mu}_{\mathrm{b}}}\big[\widehat{d}_h^{\pi}(s,a)\big]} \lesssim \frac{HSA}{K}. \tag{72}$$

**Step 4: putting everything together.** Taking (66) and (67) together with the above bounds on $\beta_1$ and $\beta_2$ (see (71) and (72)) leads to

$$\big\langle d_j^{\pi^\star}, V_j^\star - V_j^{\widehat{\pi}}\big\rangle \lesssim \beta_1\beta_2 + \frac{H^2SA\log\frac{KH}{\delta}}{K} + H^2\Big(\sqrt{\frac{SA}{HK}} + \frac{SA\xi}{N}\Big)$$

$$\lesssim \sqrt{\frac{H^3SA}{K}\log\frac{KH}{\delta}} + \frac{H^2SA\log\frac{KH}{\delta}}{K} + \frac{H^2SA\xi}{N}$$
$$\asymp \sqrt{\frac{H^3SA}{K}\log\frac{KH}{\delta}},$$

with the proviso that $K \gtrsim HSA\log\frac{KH}{\delta}$, and $N \gtrsim \sqrt{H^7S^7A^7K\log\frac{HSA}{\delta}}$. In particular, taking $j=1$ in the above display and using the fact that $d_1^{\pi^\star}(s)=\rho(s)$, we arrive at

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) = \big\langle d_1^{\pi^\star}, V_1^\star - V_1^{\widehat{\pi}}\big\rangle \lesssim \sqrt{\frac{H^3SA}{K}\log\frac{KH}{\delta}} \le \varepsilon,$$

provided that $K \ge \frac{c_K H^3SA\log\frac{KH}{\delta}}{\varepsilon^2}$ for some large enough universal constant $c_K > 0$. This concludes the proof of Theorem 3.

## 6.3 Proof of Lemma 4

We intend to establish this lemma via an inductive argument.

**Step 1: the base case with $h=1$.** Let us begin by looking at the case with $h=1$. It is seen from our construction (13) and the fact $d_1^\pi(s,a)=\rho(s)\pi_1(a\,|\,s)$ that

$$\widehat{d}_1^\pi(s,a) - d_1^\pi(s,a) = \pi_1(a\,|\,s)\bigg(\frac{1}{N}\sum_{n=1}^N \mathbb{1}(s_1^{n,0}=s) - \rho(s)\bigg). \tag{73}$$

Recall that $\{s_1^{n,0}\}_{1\le n\le N}$ are independently drawn from the initial distribution $\rho$, and hence $\{\mathbb{1}(s_1^{n,0}=s)\}_{1\le n\le N}$ are independent Bernoulli random variables with mean $\rho(s)$. Apply the Bernstein inequality (Vershynin, 2018, Theorem 2.8.4) in conjunction with the union bound to show that: there exists some universal constant $\widetilde{C}>0$ such that with probability exceeding $1-\delta$,

$$\begin{aligned}\bigg|\frac{1}{N}\sum_{n=1}^N \mathbb{1}(s_1^{n,0}=s) - \rho(s)\bigg| &\le \sqrt{\frac{\widetilde{C}}{N}\rho(s)\log\frac{SAH}{\delta}} + \frac{\widetilde{C}}{N}\log\frac{SAH}{\delta} \\ &\le \frac{1}{2}\rho(s) + \frac{\widetilde{C}}{2N}\log\frac{SAH}{\delta} + \frac{\widetilde{C}}{N}\log\frac{SAH}{\delta} \\ &\le \frac{1}{2}\rho(s) + \frac{2\widetilde{C}}{N}\log\frac{SAH}{\delta}\end{aligned} \tag{74}$$

holds simultaneously for all $s\in\mathcal{S}$, where the second line comes from the AM-GM inequality. Substituting this into (73) and recalling that $\xi = c_\xi H^3S^3A^3\log\frac{HSA}{\delta}$ give

$$\Big|\widehat{d}_1^\pi(s,a) - d_1^\pi(s,a)\Big| \le \frac{1}{2}\rho(s)\pi_1(a\,|\,s) + \frac{2\widetilde{C}\pi_1(a\,|\,s)}{N}\log\frac{SAH}{\delta} \le \frac{1}{2}d_1^\pi(s,a) + \frac{\xi}{4N}$$

as long as $c_\xi>0$ is sufficiently large, which clearly holds true for all policy $\pi$. This finishes the proof of the claim (57) when $h=1$.

**Step 2: the inductive step.** Next, we carry out the inductive step. Assuming that the claim (57) holds for the $j$-th step ($\forall 1\le j\le h$), we would like to establish its validity for the $(h+1)$-th step as well. In what follows, we introduce the following shorthand notation for the transition probabilities under policy $\pi$:

$$P_{j\to h}^\pi(s,a\,|\,s',a') = \mathbb{P}\big(s_h=s, a_h=a\,|\,s_j=s', a_j=a';\pi\big), \tag{75a}$$
$$P_h^\pi(s,a\,|\,s',a') = P_h(s\,|\,s',a')\pi_{h+1}(a\,|\,s), \tag{75b}$$
$$\widehat{P}_h^\pi(s,a\,|\,s',a') = \widehat{P}_h(s\,|\,s',a')\pi_{h+1}(a\,|\,s), \tag{75c}$$

where in (75a) the probability is calculated assuming policy $\pi$ is executed.

**Step 2.1: decomposing $\widehat{d}^{\pi}_{h+1} - d^{\pi}_{h+1}$ into two terms.** By virtue of the construction of $\widehat{d}^{\pi}_{h+1}$ in (19) as well as the basic identities $d^{\pi}_{h+1}(s) = \big\langle P_h(s\,|\,\cdot,\cdot),\, d^{\pi}_h(\cdot,\cdot)\big\rangle$ and $d^{\pi}_{h+1}(s,a) = d^{\pi}_{h+1}(s)\pi_{h+1}(a\,|\,s)$, we have

$$\begin{aligned}
\widehat{d}^{\pi}_{h+1}(s,a) - d^{\pi}_{h+1}(s,a) &= \pi_{h+1}(a\,|\,s)\Big\{\big\langle \widehat{P}_h(s\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_h(\cdot,\cdot)\big\rangle - \big\langle P_h(s\,|\,\cdot,\cdot),\, d^{\pi}_h(\cdot,\cdot)\big\rangle\Big\} \\
&= \pi_{h+1}(a\,|\,s)\Big\{\big\langle \widehat{P}_h(s\,|\,\cdot,\cdot) - P_h(s\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_h(\cdot,\cdot)\big\rangle + \big\langle P_h(s\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_h(\cdot,\cdot) - d^{\pi}_h(\cdot,\cdot)\big\rangle\Big\} \\
&= \big\langle \widehat{P}^{\pi}_h(s,a\,|\,\cdot,\cdot) - P^{\pi}_h(s,a\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_h(\cdot,\cdot)\big\rangle + \big\langle P^{\pi}_{h\to h+1}(s,a\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_h(\cdot,\cdot) - d^{\pi}_h(\cdot,\cdot)\big\rangle, \qquad (76)
\end{aligned}$$

where the last identity makes use of the notation (75). Continuing this derivation, we arrive at

$$\begin{aligned}
&\widehat{d}^{\pi}_{h+1}(s,a) - d^{\pi}_{h+1}(s,a) \\
&= \big\langle \widehat{P}^{\pi}_h(s,a\,|\,\cdot,\cdot) - P^{\pi}_h(s,a\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_h(\cdot,\cdot)\big\rangle + \sum_{s_h,a_h} P^{\pi}_{h\to h+1}(s,a\,|\,s_h,a_h)\Big(\widehat{d}^{\pi}_h(s_h,a_h) - d^{\pi}_h(s_h,a_h)\Big) \\
&= \big\langle \widehat{P}^{\pi}_h(s,a\,|\,\cdot,\cdot) - P^{\pi}_h(s,a\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_h(\cdot,\cdot)\big\rangle + \sum_{s_h,a_h} P^{\pi}_{h\to h+1}(s,a\,|\,s_h,a_h) \\
&\quad\cdot\Big\{\big\langle \widehat{P}^{\pi}_{h-1}(s_h,a_h\,|\,\cdot,\cdot) - P^{\pi}_{h-1}(s_h,a_h\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_{h-1}(\cdot,\cdot)\big\rangle + \big\langle P^{\pi}_{h-1\to h}(s_h,a_h\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_{h-1}(\cdot,\cdot) - d^{\pi}_{h-1}(\cdot,\cdot)\big\rangle\Big\} \\
&= \big\langle \widehat{P}^{\pi}_h(s,a\,|\,\cdot,\cdot) - P^{\pi}_h(s,a\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_h(\cdot,\cdot)\big\rangle \\
&\quad + \sum_{s_h,a_h} P^{\pi}_{h\to h+1}(s,a\,|\,s_h,a_h)\big\langle \widehat{P}^{\pi}_{h-1}(s_h,a_h\,|\,\cdot,\cdot) - P^{\pi}_{h-1}(s_h,a_h\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_{h-1}(\cdot,\cdot)\big\rangle \\
&\quad + \big\langle P^{\pi}_{h-1\to h+1}(s_h,a_h\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_{h-1}(\cdot,\cdot) - d^{\pi}_{h-1}(\cdot,\cdot)\big\rangle,
\end{aligned}$$

where the second identity applies (76) to the term $\widehat{d}^{\pi}_h - d^{\pi}_h$, and the last line is valid since

$$\sum_{s_h,a_h} P^{\pi}_{h\to h+1}(s,a\,|\,s_h,a_h)P^{\pi}_{h-1\to h}(s_h,a_h\,|\,s',a') = P^{\pi}_{h-1\to h+1}(s,a\,|\,s',a').$$

Repeating the above argument recursively, we arrive at

$$\begin{aligned}
&\widehat{d}^{\pi}_{h+1}(s,a) - d^{\pi}_{h+1}(s,a) \\
&= \Big\langle \widehat{P}^{\pi}_h(s,a\,|\,\cdot,\cdot) - P^{\pi}_h(s,a\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_h(\cdot,\cdot)\Big\rangle \\
&\quad + \sum_{j=1}^{h-1}\sum_{s_{j+1},a_{j+1}} P^{\pi}_{j+1\to h+1}(s,a\,|\,s_{j+1},a_{j+1})\Big\langle \widehat{P}^{\pi}_j(s_{j+1},a_{j+1}\,|\,\cdot,\cdot) - P^{\pi}_j(s_{j+1},a_{j+1}\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_j(\cdot,\cdot)\Big\rangle \\
&= \sum_{j=1}^{h}\sum_{s_{j+1},a_{j+1}} P^{\pi}_{j+1\to h+1}(s,a\,|\,s_{j+1},a_{j+1})\Big\langle \widehat{P}^{\pi}_j(s_{j+1},a_{j+1}\,|\,\cdot,\cdot) - P^{\pi}_j(s_{j+1},a_{j+1}\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_j(\cdot,\cdot)\Big\rangle. \qquad (77)
\end{aligned}$$

Next, we would like to make use of the construction of $\widehat{P}_h$ in the above decomposition (77). For this purpose, we find it convenient to introduce the following notation $k_n(s,a,h)$ such that

$$k_n(s,a,h):\ \text{the index of the episode in which the trajectory visits } (s,a) \text{ in step } h \text{ for the } n\text{-th time.} \qquad (78)$$

In view of our construction of $\widehat{P}_h$ in (16), we can divide the following inner product into two components based on whether $N_j(s_j,a_j) > \xi$ or not:

$$\begin{aligned}
&\Big\langle \widehat{P}^{\pi}_j(s_{j+1},a_{j+1}\,|\,\cdot,\cdot) - P^{\pi}_j(s_{j+1},a_{j+1}\,|\,\cdot,\cdot),\, \widehat{d}^{\pi}_j(\cdot,\cdot)\Big\rangle \\
&= \sum_{(s_j,a_j):N_j(s_j,a_j)>\xi} \frac{\widehat{d}^{\pi}_j(s_j,a_j)}{N_j(s_j,a_j)} \sum_{n=1}^{N_j(s_j,a_j)} \Big[\mathbb{1}\big(s_{j+1}^{k_n(s_j,a_j,j),j} = s_{j+1}\big) - P_j(s_{j+1}\,|\,s_j,a_j)\Big]\pi_{j+1}(a_{j+1}\,|\,s_{j+1}) \\
&\quad - \sum_{(s_j,a_j):N_j(s_j,a_j)\le\xi} P^{\pi}_j(s_{j+1},a_{j+1}\,|\,s_j,a_j)\widehat{d}^{\pi}_j(s_j,a_j).
\end{aligned}$$

Substitution into (77) allows one to decompose

$$\widehat{d}^{\pi}_{h+1}(s,a) - d^{\pi}_{h+1}(s,a) = \beta^{\pi}_{h+1}(s,a) - e^{\pi}_{h+1}(s,a), \tag{79}$$

where the two terms $e^{\pi}_{h+1}(s,a)$ and $\beta^{\pi}_{h+1}(s,a)$ are given respectively by

$$\begin{aligned} e^{\pi}_{h+1}(s,a) &\coloneqq \sum_{j=1}^{h} \sum_{s_{j+1},a_{j+1}} P^{\pi}_{j+1\to h+1}(s,a \,|\, s_{j+1},a_{j+1}) \sum_{(s_j,a_j):N_j(s_j,a_j)\le\xi} P^{\pi}_{j}(s_{j+1},a_{j+1} \,|\, s_j,a_j)\widehat{d}^{\pi}_{j}(s_j,a_j) \\ &= \sum_{j=1}^{h} \sum_{(s_j,a_j):N_j(s_j,a_j)\le\xi} P^{\pi}_{j\to h+1}(s,a \,|\, s_j,a_j)\widehat{d}^{\pi}_{j}(s_j,a_j) \ge 0 \end{aligned} \tag{80}$$

and

$$\begin{aligned} \beta^{\pi}_{h+1}(s,a) \coloneqq \sum_{j=1}^{h} \sum_{s_{j+1},a_{j+1}} P^{\pi}_{j+1\to h+1}(s,a \,|\, s_{j+1},a_{j+1}) \sum_{(s_j,a_j):N_j(s_j,a_j)>\xi} \frac{\widehat{d}^{\pi}_{j}(s_j,a_j)}{N_j(s_j,a_j)} \\ \cdot \sum_{n=1}^{N_j(s_j,a_j)} \Big[\mathbb{1}\big(s_{j+1}^{k_n(s_j,a_j,j),j} = s_{j+1}\big) - P_j(s_{j+1} \,|\, s_j,a_j)\Big]\pi_{j+1}(a_{j+1} \,|\, s_{j+1}). \end{aligned} \tag{81}$$

In words, $\{e^{\pi}_{h+1}(s,a)\}$ captures the total contributions of those state-action pairs $(s_j,a_j)$ with $N_j(s_j,a_j)\le\xi$, while $\beta^{\pi}_{h+1}(s,a)$ reflects the contributions of the remaining state-action pairs. In order to establish Lemma 4, it boils down to controlling $e^{\pi}_{h+1}(s,a)$ and $\beta^{\pi}_{h+1}(s,a)$ in the decomposition (79), which we shall accomplish separately in the following.

**Step 2.2: controlling the term $\sum_{s,a} e^{\pi}_{h+1}(s,a)$.** In order to bound the sum $\sum_{s,a} e^{\pi}_{h+1}(s,a)$ (cf. (80)), we first make the following claim: with probability exceeding $1-\delta$,

$$\Big\{(s,a): N_j(s,a)\le\xi\Big\} \overset{\text{(a)}}{\subseteq} \left\{(s,a): \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\big[d^{\pi}_j(s,a)\big] \le \frac{3\xi}{N}\right\} \overset{\text{(b)}}{\subseteq} \left\{(s,a): \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\big[\widehat{d}^{\pi}_j(s,a)\big] \le \frac{6.5\xi}{N}\right\} =: \mathcal{J}_j \tag{82}$$

holds for all $1\le j\le h$, which we shall justify below.

- *Inclusion relation (a).* To see why the first relation (a) is valid, recall that $N_j(s,a)$ can be viewed as the sum of $N$ independent Bernoulli random variables and obeys

$$\mathbb{E}\big[N_j(s,a)\big] = N \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\big[d^{\pi}_j(s,a)\big],$$

given that in this round we take samples using the exploration policy $\pi^{\mathsf{explore},j} = \mathbb{E}_{\pi\sim\widehat{\mu}^j}[\pi]$ (see line 9 of Algorithm 2). Repeating the Bernstein-type concentration argument as in (74) and applying the union bound, we see that with probability at least $1-\delta$,

$$N_j(s,a) \ge \frac{N}{2} \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\big[d^{\pi}_j(s,a)\big] - 2\widetilde{C}\log\frac{HSA}{\delta} \tag{83}$$

holds simultaneously for all $1\le j\le h$ and all $(s,a)\in\mathcal{S}\times\mathcal{A}$, where $\widetilde{C}>0$ is some universal constant. Consequently, for any $(s,a)$ obeying $N_j(s,a)\le\xi$ one has

$$\mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\big[d^{\pi}_j(s,a)\big] \le \frac{2N_j(s,a)}{N} + \frac{4\widetilde{C}}{N}\log\frac{HSA}{\delta} \le \frac{3\xi}{N},$$

where the last inequality also relies on the choice (18) of $\xi$. This establishes the advertised relation (a).

- *Inclusion relation (b).* Regarding the second claim (b) in (82), one can easily verify it using the induction hypothesis (57) for any $j \le h$; namely, for any $(s,a)$ obeying $\mathbb{E}_{\pi\sim\widehat{\mu}^j}\left[d_j^\pi(s,a)\right] \le \frac{3\xi}{N}$, one has

$$\begin{aligned} \frac{1}{2}\mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[\widehat{d}_j^\pi(s,a)\right] - \frac{\xi}{4N} &\le \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[d_j^\pi(s,a)\right] \le \frac{3\xi}{N} \\ \Longrightarrow \qquad \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[\widehat{d}_j^\pi(s,a)\right] &\le \frac{6.5\xi}{N} \end{aligned} \tag{84}$$

Thus far, we have validated the claim (82). With this result in place, invoke the definition (80) to derive

$$\begin{aligned} \sum_{s,a} e_{h+1}^\pi(s,a) &= \sum_{s,a}\sum_{j=1}^{h}\sum_{(s_j,a_j):N_j(s_j,a_j)\le\xi} P_{j\to h+1}^\pi(s,a\,|\,s_j,a_j)\widehat{d}_j^\pi(s_j,a_j) \\ &= \sum_{j=1}^{h}\sum_{(s_j,a_j):N_j(s_j,a_j)\le\xi}\left(\sum_{s,a} P_{j\to h+1}^\pi(s,a\,|\,s_j,a_j)\right)\widehat{d}_j^\pi(s_j,a_j) \\ &\overset{\text{(i)}}{\le} \sum_{j=1}^{h}\sum_{(s_j,a_j)\in\mathcal{J}_j}\widehat{d}_j^\pi(s_j,a_j) \\ &\overset{\text{(ii)}}{\le} \sum_{j=1}^{h}\sum_{(s_j,a_j)\in\mathcal{J}_j}\widehat{d}_j^\pi(s,a)\cdot\frac{\frac{1}{KH}+\frac{6.5\xi}{N}}{\frac{1}{KH}+\mathbb{E}_{\pi'\sim\widehat{\mu}^j}[\widehat{d}_h^{\pi'}(s,a)]} \\ &\le \left(\frac{1}{KH}+\frac{6.5\xi}{N}\right)\sum_{j=1}^{h}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\frac{1}{KH}+\widehat{d}_j^\pi(s,a)}{\frac{1}{KH}+\mathbb{E}_{\pi'\sim\widehat{\mu}^j}[\widehat{d}_h^{\pi'}(s,a)]}. \end{aligned} \tag{85}$$

Here, (i) arises from (82) and the fact $\sum_{s,a} P_{j\to h+1}^\pi(s,a\,|\,s_j,a_j) = 1$, whereas (ii) is a consequence of the definition of $\mathcal{J}_j$ in (82). Moreover, suppose that the subroutine specified in Algorithm 2 for step $j$ terminates with iterates $\widehat{\mu}^j = \mu^{(t)}$ and $\pi^{(t)}$, where we recall that $\pi^{(t)}$ is a deterministic Markov policy chosen to obey

$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\widehat{d}_h^{\pi^{(t)}}(s,a)}{\frac{1}{KH}+\mathbb{E}_{\pi'\sim\widehat{\mu}^j}\left[\widehat{d}_h^{\pi'}(s,a)\right]} = \max_{\pi\in\Pi}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\widehat{d}_h^{\pi}(s,a)}{\frac{1}{KH}+\mathbb{E}_{\pi'\sim\widehat{\mu}^j}\left[\widehat{d}_h^{\pi'}(s,a)\right]}. \tag{86}$$

The stopping criterion specified in line 6 of Algorithm 2 then reveals that

$$2SA \ge g\big(\pi^{(t)},\widehat{d},\widehat{\mu}^j\big) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\frac{1}{KH}+\widehat{d}_h^{\pi^{(t)}}(s,a)}{\frac{1}{KH}+\mathbb{E}_{\pi'\sim\widehat{\mu}^j}\left[\widehat{d}_h^{\pi'}(s,a)\right]} = \max_{\pi\in\Pi}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\frac{1}{KH}+\widehat{d}_h^{\pi}(s,a)}{\frac{1}{KH}+\mathbb{E}_{\pi'\sim\widehat{\mu}^j}\left[\widehat{d}_h^{\pi'}(s,a)\right]}. \tag{87}$$

Substitution into (85) indicates that

$$\begin{aligned} \sum_{s,a} e_{h+1}^\pi(s,a) &\le \left(\frac{1}{KH}+\frac{6.5\xi}{N}\right)2SAH = \frac{2SA}{K}+\frac{13c_\xi S^4A^4H^4\log\frac{HSA}{\delta}}{N} \\ &\lesssim \sqrt{\frac{SA}{HK}}, \end{aligned} \tag{88}$$

where the validity of the last line is guaranteed as long as $K \gtrsim HSA$ and $N \gtrsim \sqrt{H^9S^7A^7K}\log\frac{HSA}{\delta}$.

**Step 2.3: controlling the term $\beta_{h+1}^\pi(s,a)$.** Next, we turn attention to controlling $\beta_{h+1}^\pi(s,a)$. Recall from (81) that

$$\beta_{h+1}^\pi(s,a) = \sum_{j=1}^{h}\sum_{(s_j,a_j):N_j(s_j,a_j)>\xi}\sum_{n=1}^{N_j(s_j,a_j)}$$

$$\underbrace{\frac{\widehat{d}_j^{\pi}(s_j,a_j)}{N_j(s_j,a_j)}\sum_{s_{j+1},a_{j+1}} P^{\pi}_{j+1\to h+1}(s,a\,|\,s_{j+1},a_{j+1})\Big[\mathbb{1}\left(s_{j+1}^{k_n(s_j,a_j),j}=s_{j+1}\right)-P_j(s_{j+1}\,|\,s_j,a_j)\Big]\pi_{j+1}(a_{j+1}\,|\,s_{j+1})}_{=:X_j^n(s_j,a_j)},$$

which we shall bound by resorting to the Bernstein inequality. Consider each $1\le j\le h$. Evidently, when conditional on $\{N_j(s_j,a_j):(s_j,a_j)\in\mathcal{S}\times\mathcal{A}\}$, the random variables $\{X_j^n(s_j,a_j):(s_j,a_j)\in\mathcal{S}\times\mathcal{A}, 1\le n\le N_j(s_j,a_j)\}$ are statistically independent with mean zero. Let us look at the size of each term as well as the variance statistics separately, both of which are needed in order to apply the Bernstein inequality.

*Step 2.3.1: bounding the size of each $X_j^n(s_j,a_j)$.* Towards this end, we make the following two observations:

- Firstly, let us write

$$\begin{aligned}&\sum_{s_{j+1},a_{j+1}} P^{\pi}_{j+1\to h+1}(s,a\,|\,s_{j+1},a_{j+1})\,\mathbb{1}\left(s_{j+1}^{k_n(s_j,a_j),j+1}=s_{j+1}\right)\pi_{j+1}(a_{j+1}\,|\,s_{j+1})\\ &\quad=\sum_{s_{j+1}}\mathbb{1}\left(s_{j+1}^{k_n(s_j,a_j),j+1}=s_{j+1}\right)\underbrace{\sum_{a_{j+1}} P^{\pi}_{j+1\to h+1}(s,a\,|\,s_{j+1},a_{j+1})\pi_{j+1}(a_{j+1}\,|\,s_{j+1})}_{=:Y(s_{j+1},s,a)},\end{aligned}\tag{89}$$

  where $Y(s_{j+1},s,a)$ is the transition probability from state $s_{j+1}$ in the $(j+1)$-th step to the state-action pair $(s,a)$ in the $(h+1)$-th step under policy $\pi$. Hence, the sum in (89) is bounded above by 1.

- Secondly, it is seen that

$$\sum_{s_{j+1},a_{j+1}} P^{\pi}_{j+1\to h+1}(s,a\,|\,s_{j+1},a_{j+1})P_j(s_{j+1}\,|\,s_j,a_j)\pi_{j+1}(a_{j+1}\,|\,s_{j+1})=P^{\pi}_{j\to h+1}(s,a\,|\,s_j,a_j)$$

  is the transition probability from the state-action pair $(s_j,a_j)$ in the $j$-th step to $(s,a)$ in the $(h+1)$-th step under policy $\pi$, which is again bounded above by 1.

- The preceding two observations taken collectively with the definition of $X_j^n(s_j,a_j)$ imply that: for any state-action pair $(s_j,a_j)$ such that $N_j(s_j,a_j)>\xi$, we have

$$\left|X_j^n(s_j,a_j)\right|\le 2\frac{\widehat{d}_j^{\pi}(s_j,a_j)}{N_j(s_j,a_j)}.\tag{90}$$

While (90) already offers an upper bound on the size of $X_j^n(s_j,a_j)$, we can further bound it by exploiting other properties about $N_j(s_j,a_j)$, given that we have restricted attention to those state-action pairs obeying $N_j(s_j,a_j)>\xi$. In view of Bernstein's inequality (Vershynin, 2018, Theorem 2.8.4) and the sampling policy $\pi^{\mathsf{explore},j}$ in use (see line 9 of Algorithm 2), there exists some universal constant $\widetilde{C}>0$ such that with probability exceeding $1-\delta$,

$$\left|N_j(s_j,a_j)-N\mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[d_j^{\pi}(s_j,a_j)\right]\right|\le\widetilde{C}\sqrt{N\mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[d_j^{\pi}(s_j,a_j)\right]\log\frac{HSA}{\delta}}+\widetilde{C}\log\frac{HSA}{\delta}\tag{91}$$

holds for any $1\le j\le h$ and $(s_j,a_j)\in\mathcal{S}\times\mathcal{A}$. Therefore, for any $(s_j,a_j)$ such that $N_j(s_j,a_j)>\xi$, we have

$$\xi<N_j(s_j,a_j)\le N\mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[d_j^{\pi}(s_j,a_j)\right]+\widetilde{C}\sqrt{N\mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[d_j^{\pi}(s_j,a_j)\right]\log\frac{HSA}{\delta}}+\widetilde{C}\log\frac{HSA}{\delta},$$

which together with the choice $\xi=c_\xi H^3S^3A^3\log\frac{HSA}{\delta}$ indicates that

$$N\mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[d_j^{\pi}(s_j,a_j)\right]\ge\frac{3}{4}\xi=\frac{3c_\xi}{4}H^3S^3A^3\log\frac{HSA}{\delta}\tag{92}$$

as long as $c_\xi > 0$ is sufficiently large. This combined with (91) gives

$$\begin{aligned} N_j(s_j,a_j) &\geq N \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[d_j^\pi(s_j,a_j)\right] - \widetilde{C}\sqrt{N \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[d_j^\pi(s_j,a_j)\right]\log\frac{HSA}{\delta}} - \widetilde{C}\log\frac{HSA}{\delta} \\ &\geq \frac{1}{2}N \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[d_j^\pi(s_j,a_j)\right] \geq \frac{1}{4}N \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[d_j^\pi(s_j,a_j)\right] + \frac{3}{16}\xi \\ &\geq \frac{1}{8}N \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[\widehat{d}_j^\pi(s_j,a_j)\right] - \frac{1}{16}\xi + \frac{3}{16}\xi \\ &\geq \frac{1}{8}\Big(N \mathop{\mathbb{E}}_{\pi\sim\widehat{\mu}^j}\left[\widehat{d}_j^\pi(s_j,a_j)\right] + 1\Big), \end{aligned} \tag{93}$$

where the second line makes use of (92), and the third line invokes the induction hypothesis (57) for the $j$-th step. Therefore, we can deduce that

$$\sum_{(s_j,a_j):N_j(s_j,a_j)>\xi} \frac{\widehat{d}_j^\pi(s_j,a_j)}{N_j(s_j,a_j)} \leq \sum_{(s_j,a_j):N_j(s_j,a_j)>\xi} \frac{8}{N}\cdot\frac{\widehat{d}_j^\pi(s_j,a_j)}{1/N + \mathbb{E}_{\pi'\sim\widehat{\mu}^j}\left[\widehat{d}_j^{\pi'}(s_j,a_j)\right]} \leq \frac{16SA}{N}, \tag{94}$$

where the first relation follows from (93), and the second relation follows from the assumption $N \leq KH$ and inequality (87). Combine (90) and (94) to yield

$$L_j \coloneqq \max_{(s_j,a_j):N_j(s_j,a_j)>\xi} \max_{1\leq n\leq N_j(s_j,a_j)} |X_j^n(s_j,a_j)| \leq \frac{16SA}{N}. \tag{95}$$

*Step 2.3.2: controlling the variance statistics.* Consider now any given deterministic Markov policy $\pi \in \Pi$. Conditional on $\{N_j(s_j,a_j) : (s_j,a_j) \in \mathcal{S}\times\mathcal{A}\}$ and $\{\widehat{d}_j^\pi(s_j,a_j) : (s_j,a_j) \in \mathcal{S}\times\mathcal{A}\}$, we can calculate that

$$\begin{aligned} \mathsf{Var}\big(X_j^n(s_j,a_j)\big) &\leq \left(\frac{\widehat{d}_j^\pi(s_j,a_j)}{N_j(s_j,a_j)}\right)^2 \sum_{s_{j+1},a_{j+1}} P_j(s_{j+1}\,|\,s_j,a_j)\Big[P_{j+1\to h+1}^\pi(s,a\,|\,s_{j+1},a_{j+1})\pi_{j+1}(a_{j+1}\,|\,s_{j+1})\Big]^2 \\ &\leq \left(\frac{\widehat{d}_j^\pi(s_j,a_j)}{N_j(s_j,a_j)}\right)^2 \sum_{s_{j+1},a_{j+1}} P_{j+1\to h+1}^\pi(s,a\,|\,s_{j+1},a_{j+1})P_j^\pi(s_{j+1},a_{j+1}\,|\,s_j,a_j) \\ &= \left(\frac{\widehat{d}_j^\pi(s_j,a_j)}{N_j(s_j,a_j)}\right)^2 P_{j\to h+1}^\pi(s,a\,|\,s_j,a_j), \end{aligned}$$

where the second line is valid since

$$P_j(s_{j+1}\,|\,s_j,a_j)\pi_{j+1}(a_{j+1}\,|\,s_{j+1})P_{j+1\to h+1}^\pi(s,a\,|\,s_{j+1},a_{j+1}) = P_{j+1\to h+1}^\pi(s,a\,|\,s_{j+1},a_{j+1})P_j^\pi(s_{j+1},a_{j+1}\,|\,s_j,a_j).$$

As a consequence, conditional on $\{N_j(s_j,a_j) : (s_j,a_j) \in \mathcal{S}\times\mathcal{A}\}$ and $\{\widehat{d}_j^\pi(s_j,a_j) : (s_j,a_j) \in \mathcal{S}\times\mathcal{A}\}$ one has

$$\begin{aligned} V_j &\coloneqq \sum_{(s_j,a_j):N_j(s_j,a_j)>\xi} \sum_{n=1}^{N_j(s_j,a_j)} \mathsf{Var}\big(X_j^n(s_j,a_j)\big) \\ &\leq \sum_{(s_j,a_j):N_j(s_j,a_j)>\xi} \frac{\left[\widehat{d}_j^\pi(s_j,a_j)\right]^2}{N_j(s_j,a_j)} P_{j\to h+1}^\pi(s,a\,|\,s_j,a_j) \\ &\leq \left(\max_{(s_j,a_j):N_j(s_j,a_j)>\xi} \frac{\widehat{d}_j^\pi(s_j,a_j)}{N_j(s_j,a_j)}\right)\left(\sum_{s_j,a_j} \widehat{d}_j^\pi(s_j,a_j)P_{j\to h+1}^\pi(s,a\,|\,s_j,a_j)\right) \\ &\leq \frac{16SA}{N}\Big(2d_{h+1}^\pi(s,a) + \frac{SA\xi}{2N}\Big). \end{aligned} \tag{96}$$

Here, the last inequality follows from (94) and

$$\sum_{s_j,a_j} \widehat{d}_j^{\pi}(s_j,a_j) P_{j\to h+1}^{\pi}(s,a\,|\,s_j,a_j) \le 2\sum_{s_j,a_j}\left(d_j^{\pi}(s_j,a_j)+\frac{\xi}{4N}\right) P_{j\to h+1}^{\pi}(s,a\,|\,s_j,a_j)$$
$$\le 2d_{h+1}^{\pi}(s,a)+\frac{SA\xi}{2N},$$

where the first line invokes the induction hypothesis (57) for the $j$-th step.

*Step 2.3.3: invoking the Bernstein inequality and union bound.* Armed with the above results, we now develop some concentration bounds for $\beta_{h+1}^{\pi}(s,a)$. Towards this, let us begin by looking at any given deterministic Markov policy $\pi\in\Pi$. By virtue of the Bernstein inequality, there exist some universal constants $C_1,C_2>0$ such that: for any $1\le j\le h$ and any given $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\left|\sum_{(s_j,a_j):N_j(s_j,a_j)>\xi}\sum_{n=1}^{N_j(s_j,a_j)} X_j^n(s_j,a_j)\right| \le C_1\left[\sqrt{V_j\log\frac{|\Pi|SAH}{\delta}}+L_j\log\frac{|\Pi|SAH}{\delta}\right]$$
$$\le C_2\sqrt{\frac{HS^2A}{N}\left(d_{h+1}^{\pi}(s,a)+\frac{SA\xi}{N}\right)\log\frac{SAH}{\delta}}+C_2\frac{HS^2A}{N}\log\frac{SAH}{\delta}$$
$$\le C_2\sqrt{\frac{HS^2A}{N}d_{h+1}^{\pi}(s,a)\log\frac{SAH}{\delta}}+C_2\sqrt{\frac{HS^3A^2\xi}{N^2}\log\frac{SAH}{\delta}}+C_2\frac{HS^2A}{N}\log\frac{SAH}{\delta} \tag{97}$$

holds true with probability at least $1-\delta/(|\Pi|HSA)$, where the second line relies on (95), (96) as well as the fact that $|\Pi|\le A^{HS}$, and the third line invokes the elementary inequality $\sqrt{x+y}\le\sqrt{x}+\sqrt{y}$. Taking the union bound over all $(s,a,j)\in\mathcal{S}\times\mathcal{A}\times[h]$ and substituting (97) into the expression $\beta_{h+1}^{\pi}(s,a)$ yield: with probability exceeding $1-\delta$,

$$\left|\beta_{h+1}^{\pi}(s,a)\right| \le C_2H\sqrt{\frac{HS^2A}{N}d_{h+1}^{\pi}(s,a)\log\frac{SAH}{\delta}}+C_2H\sqrt{\frac{HS^3A^2\xi}{N^2}\log\frac{SAH}{\delta}}+C_2\frac{H^2S^2A}{N}\log\frac{SAH}{\delta}$$
$$\le \frac{1}{2}d_{h+1}^{\pi}(s,a)+\frac{1}{2}C_2^2\frac{H^3S^2A}{N}\log\frac{SAH}{\delta}+\frac{\xi}{8N}+2C_2^2\frac{H^3S^3A^2}{N}\log\frac{SAH}{\delta}+C_2\frac{H^2S^2A}{N}\log\frac{SAH}{\delta}$$
$$\le \frac{1}{2}d_{h+1}^{\pi}(s,a)+\frac{\xi}{4N}$$

holds simultaneously for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and all deterministic Markov policies $\pi\in\Pi$. Here, the penultimate relation follows from the AM-GM inequality, whereas the last inequality holds as long as $c_\xi>0$ is sufficiently large. Taking the union bound over all deterministic policies $\pi\in\Pi$, we arrive at

$$\left|\beta_{h+1}^{\pi}(s,a)\right| \le \frac{1}{2}d_{h+1}^{\pi}(s,a)+\frac{\xi}{4N} \qquad \text{for all } \pi\in\Pi \text{ and } (s,a)\in\mathcal{S}\times\mathcal{A}. \tag{98}$$

**Step 2.3: putting everything together.** Substituting (88) and (98) into (79), we immediately establish the claim (57) for step $h+1$. This together with standard induction arguments concludes the proof of Lemma 4.

# 7 Analysis for reward-free exploration

We now turn attention to the proof of Theorem 2, which is concerned with the reward-free setting that seeks to cover all possible reward functions simultaneously. The proof of Theorem 2 largely resembles that of Theorem 1, except that we are in need of a different uniform concentration result as follows.

**Lemma 5.** *Let $\widehat{P}=\{\widehat{P}_h\}_{1\le h\le H}$ denote the empirical transition kernel constructed in Algorithm 4. With probability at least $1-\delta$, one has*

$$\left|\left(\widehat{P}_{h,s,a}-P_{h,s,a}\right)V\right| \le \sqrt{\frac{48S}{\widehat{N}_h^{\flat}(s,a)}\mathsf{Var}_{\widehat{P}_{h,s,a}}(V)\log\frac{KH}{\delta}}+\frac{64HS}{\widehat{N}_h^{\flat}(s,a)}\log\frac{KH}{\delta} \tag{99a}$$

$$\mathsf{Var}_{\widehat{P}_{h,s,a}} \le 8\mathsf{Var}_{P_{h,s,a}}(V) + \frac{10H^2S}{\widehat{N}_h^{\mathsf{b}}(s,a)} \log \frac{KH}{\delta} \tag{99b}$$

*hold simultaneously for all* $V \in \mathbb{R}^S$ *obeying* $\|V\|_\infty \le H$ *and all* $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

In words, this lemma extends the result in Lemma 3 to provide uniform control over all possible vectors $V$. The proof of this lemma is provided in Section 7.1.

Armed with Lemma 5, we can repeat the same analysis as in Li et al. (2022b, Section 7) to demonstrate that: with probability exceeding $1-\delta$,

$$\left\langle d_h^{\pi^\star}, V_h^\star - V_h^{\widehat{\pi}} \right\rangle \le 2 \sum_{j:j\ge h} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_j^{\pi^\star}(s,a) b_j(s,a) \tag{100}$$

holds simultaneously for all possible reward functions, where the penalty function $b_j(s,a)$ is defined in (106). As a result, one can then repeat the same analysis as in the proof of Theorem 1 in Section 6.2 to establish the desired sample complexity for the reward-free case. The only additional thing that needs to be taken care of now is that we should replace the penalty function $b_j(s,a)$ defined in (105) for the reward-agnostic counterpart with (106) for the reward-free case. This in turn leads to

$$\left\langle d_j^{\pi^\star}, V_j^\star - V_j^{\widehat{\pi}} \right\rangle \lesssim \sqrt{\frac{H^3S^2A}{K}\log\frac{KH}{\delta}} + \frac{H^2S^2A\log\frac{KH}{\delta}}{K} + H^2\left(\sqrt{\frac{SA}{HK}} + \frac{SA\xi}{N}\right) \asymp \sqrt{\frac{H^3S^2A}{K}\log\frac{KH}{\delta}}$$

as long as $KH \ge N \gtrsim \sqrt{H^9S^7A^7K}\log\frac{HSA}{\delta}$, thus indicating that

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) = \left\langle d_1^{\pi^\star}, V_1^\star - V_1^{\widehat{\pi}} \right\rangle \lesssim \sqrt{\frac{H^3S^2A}{K}\log\frac{KH}{\delta}} \le \varepsilon,$$

as long as $K \ge \frac{c_K H^3S^2A\log\frac{KH}{\delta}}{\varepsilon^2}$ for some sufficiently large constant $c_K > 0$. We omite other details for the sake of brevity.

## 7.1 Proof of Lemma 5

Let us construct an $\epsilon$-net $\mathcal{N}$ for the set $[0,H]^S$ under metric $\|\cdot\|_\infty$; as is well known, one can choose $\mathcal{N}$ such that $|\mathcal{N} \le |(H/\epsilon)^S$ (Vershynin, 2018). In view of Li et al. (2022b, Lemma 8), we know that any vector $\widetilde{V} \in \mathcal{N}$ satisfies

$$\left|(\widehat{P}_{h,s,a} - P_{h,s,a})\widetilde{V}\right| \le \sqrt{\frac{48}{\widehat{N}_h^{\mathsf{b}}(s,a)}\mathsf{Var}_{\widehat{P}_{h,s,a}}(\widetilde{V})\log\frac{KH|\mathcal{N}|}{\delta}} + \frac{48H}{\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH|\mathcal{N}|}{\delta}$$

and

$$\mathsf{Var}_{\widehat{P}_{h,s,a}}(\widetilde{V}) \le 2\mathsf{Var}_{P_{h,s,a}}(\widetilde{V}) + \frac{5H^2}{3\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH|\mathcal{N}|}{\delta}$$

with probability exceeding $1-\delta/|\mathcal{N}|$. Taking the union bound over all $\widetilde{V} \in \mathcal{N}$, we know that with probability exceeding $1-\delta$,

$$\left|(\widehat{P}_{h,s,a} - P_{h,s,a})\widetilde{V}\right| \le \sqrt{\frac{48S}{\widehat{N}_h^{\mathsf{b}}(s,a)}\mathsf{Var}_{\widehat{P}_{h,s,a}}(\widetilde{V})\log\frac{KH^2}{\delta\epsilon}} + \frac{48HS}{\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\epsilon}$$

and

$$\mathsf{Var}_{\widehat{P}_{h,s,a}}(\widetilde{V}) \le 2\mathsf{Var}_{P_{h,s,a}}(\widetilde{V}) + \frac{5H^2S}{3\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\epsilon} \tag{101}$$

hold simultaneously for all $\widetilde{V} \in \mathcal{N}$.

We now look at an arbitrary $V \in \mathbb{R}^S$ obeying $\|V\|_\infty \le H$. From the definition of the $\epsilon$-net, we know the existence of some $\widetilde{V} \in \mathcal{N}$ such that $\|V - \widetilde{V}\|_\infty \le \epsilon$. By choosing $\epsilon = 1/K$, we can deduce that

$$
\begin{aligned}
\left|(\widehat{P}_{h,s,a} - P_{h,s,a})V\right| &\le \left|(\widehat{P}_{h,s,a} - P_{h,s,a})\widetilde{V}\right| + 2\left\|\widetilde{V} - V\right\|_\infty \\
&\le \sqrt{\frac{48S}{\widehat{N}_h^{\mathsf{b}}(s,a)}\mathsf{Var}_{\widehat{P}_{h,s,a}}(\widetilde{V})\log\frac{KH^2}{\delta\epsilon}} + \frac{48HS}{\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\epsilon} + 2\epsilon \\
&\le \sqrt{\frac{96S}{\widehat{N}_h^{\mathsf{b}}(s,a)}\mathsf{Var}_{\widehat{P}_{h,s,a}}(V)\log\frac{KH^2}{\delta\epsilon}} + \sqrt{\frac{96S}{\widehat{N}_h^{\mathsf{b}}(s,a)}\mathsf{Var}_{\widehat{P}_{h,s,a}}(V-\widetilde{V})\log\frac{KH^2}{\delta\epsilon}} + \frac{48HS}{\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\epsilon} + 2\epsilon \\
&\le \sqrt{\frac{96S}{\widehat{N}_h^{\mathsf{b}}(s,a)}\mathsf{Var}_{\widehat{P}_{h,s,a}}(V)\log\frac{KH^2}{\delta\epsilon}} + \sqrt{\frac{96S\epsilon^2}{\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\epsilon}} + \frac{48HS}{\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\epsilon} + 2\epsilon \\
&\le \sqrt{\frac{48S}{\widehat{N}_h^{\mathsf{b}}(s,a)}\mathsf{Var}_{\widehat{P}_{h,s,a}}(\widetilde{V})\log\frac{KH^2}{\delta\epsilon}} + \frac{60HS}{\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\epsilon} + 2\epsilon^2 + 2\epsilon \\
&\le \sqrt{\frac{48S}{\widehat{N}_h^{\mathsf{b}}(s,a)}\mathsf{Var}_{\widehat{P}_{h,s,a}}(\widetilde{V})\log\frac{KH^2}{\delta\epsilon}} + \frac{64HS}{\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\epsilon},
\end{aligned}
$$

where the third line is valid since $\mathsf{Var}(X+Y) \le 2\mathsf{Var}(X) + 2\mathsf{Var}(Y)$, the penultimate line follows from the AM-GM inequality, and the last inequality holds since $\epsilon = 1/K \le 1/\widehat{N}_h^{\mathsf{b}}(s,a)$. In addition, one also obtains

$$
\begin{aligned}
\mathsf{Var}_{\widehat{P}_{h,s,a}}(V) &\le 2\mathsf{Var}_{\widehat{P}_{h,s,a}}(\widetilde{V}) + 2\mathsf{Var}_{\widehat{P}_{h,s,a}}(V-\widetilde{V}) \le 2\mathsf{Var}_{\widehat{P}_{h,s,a}}(\widetilde{V}) + 2\left\|\widetilde{V} - V\right\|_\infty^2 \\
&\le 4\mathsf{Var}_{P_{h,s,a}}(\widetilde{V}) + \frac{10H^2S}{3\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\varepsilon} + 2\left\|\widetilde{V} - V\right\|_\infty^2 \\
&\le 8\mathsf{Var}_{P_{h,s,a}}(V) + 8\mathsf{Var}_{P_{h,s,a}}(V-\widetilde{V}) + \frac{10H^2S}{3\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\varepsilon} + 2\left\|\widetilde{V} - V\right\|_\infty^2 \\
&\le 8\mathsf{Var}_{P_{h,s,a}}(V) + \frac{10H^2S}{3\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\varepsilon} + 10\left\|\widetilde{V} - V\right\|_\infty^2 \\
&\le 8\mathsf{Var}_{P_{h,s,a}}(V) + \frac{10H^2S}{3\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\varepsilon} + 10\epsilon^2 \\
&\le 8\mathsf{Var}_{P_{h,s,a}}(V) + \frac{10H^2S}{\widehat{N}_h^{\mathsf{b}}(s,a)}\log\frac{KH^2}{\delta\varepsilon},
\end{aligned}
$$

where the first and the third line arise since $\mathsf{Var}(X+Y) \le 2\mathsf{Var}(X) + 2\mathsf{Var}(Y)$, and the second line comes from (101). This concludes the proof of this lemma.

# 8 Discussion

In this paper, we have introduced a reward-agnostic pure exploration algorithm that works statistically optimally for two scenarios: (a) the case where there are at most polynomially many reward functions of interest; (b) the case where one needs to accommodate an arbitrarily large number of unknown reward functions. Drawing on insights from recent advances in offline reinforcement learning, our algorithm design differs drastically from prior reward-agnostic/reward-free algorithms, and illuminates new connections between online and offline reinforcement learning.

We conclude this paper by pointing out a few directions worthy of future investigation. To begin with, the current algorithm is only guaranteed to work when the target accuracy level $\varepsilon$ is small enough; put another way, this means that our algorithm might incur a high burn-in cost, so that its optimality does not come into effect until the total sample size exceeds the burn-in cost. Can we hope to improve the algorithm design so as to cover the full $\varepsilon$ range? Additionally, how to extend the ideas of the current algorithm to

accommodate the case where low-dimensional representation of the MDP is available, in the hope of further enhancing data efficiency? Furthermore, the notion of minimax optimality might be too conservative in some practical applications. It would be of interest to design more adaptive exploration paradigms that achieve some sort of instance optimality, if not infeasible.


## Acknowledgements

Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grant FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009 and CCF-1907661. Y. Yan is supported in part by Charlotte Elizabeth Procter Honorific Fellowship from Princeton University. J. Fan's research was partially supported by the NSF grants DMS-2210833 and ONR grant N00014-22-1-2340.


## A A pessimistic model-based algorithm for offline RL

In this section, we present the precise procedure for the model-based offline RL algorithm studied in Li et al. (2022b). Before proceeding, we first convert the $K$ sample episodes collected in Stage 1.2 into a dataset of the following form:

$$\mathcal{D} = \big\{ \big(s_h^{n,\mathsf{b}}, a_h^{n,\mathsf{b}}, s_{h+1}^{n,\mathsf{b}}\big) \big\}_{1\le n\le K, 1\le h<H}, \tag{102}$$

comprising all sample transitions in these $K$ episodes. In particular, for each $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$ we define

$$N_h(s,a) \coloneqq \sum_{n=1}^{K} \mathbb{1}\big\{(s_h^{n,\mathsf{b}}, a_h^{n,\mathsf{b}}) = (s,a)\big\}, \tag{103}$$

which stands for the total number of sample transitions from the state-action pair $(s,a)$ at the $h$-th step.

**Empirical MDP.** Given the data set $\mathcal{D}$ in (102), we first subsample $\mathcal{D}$ to obtain another dataset $\mathcal{D}^{\mathsf{trim}}$ such that for each $(s,a,h)$, $\mathcal{D}^{\mathsf{trim}}$ contains exactly $\min\big\{\widehat{N}_h^{\mathsf{b}}(s,a), N_h(s,a)\big\}$ sample transitions from $(s,a)$ at step $h$. Here, we remind the reader of the definition of $\widehat{N}_h^{\mathsf{b}}(s,a)$ in (22). We can then compute the empirical transition kernel $\widehat{P} = \{\widehat{P}_h\}_{1\le h\le H}$ as follows:

$$\widehat{P}_h(s' \,|\, s,a) = \frac{\mathbb{1}\big(\min\big\{\widehat{N}_h^{\mathsf{b}}(s,a), N_h(s,a)\big\} > 0\big)}{\min\big\{\widehat{N}_h^{\mathsf{b}}(s,a), N_h(s,a)\big\}} \sum_{(s_i,a_i,h_i,s_i')\in\mathcal{D}^{\mathsf{trim}}} \mathbb{1}\big\{(s_i,a_i,h_i,s_i') = (s,a,h,s')\big\} \tag{104}$$

for each $(s,a,h,s')\in\mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{S}$. Here, we abuse the notation $\widehat{P}$ as long as it is clear from the context.

**Pessimism in the face of uncertainty.** An effective strategy to solve offline RL is to resort to the pessimism principle in the face of uncertainty. Towards this end, one needs to specify how to quantify the uncertainty of value estimation, which specify now. In our algorithm, we choose the penalty term $b_h(s,a)$ for each $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$ based on Bernstein-style concentration bounds; more specifically,

- In the reward-agnostic setting, the Bernstein-style penalty is chosen to be

$$b_h(s,a) = \min\left\{ \sqrt{\frac{c_{\mathsf{b}}\log\frac{HSA}{\delta}}{\min\big\{\widehat{N}_h^{\mathsf{b}}(s,a), N_h(s,a)\big\}} \mathsf{Var}_{\widehat{P}_{h,s,a}}\big(\widehat{V}_{h+1}\big)} + c_{\mathsf{b}} H \frac{\log\frac{HSA}{\delta}}{\min\big\{\widehat{N}_h^{\mathsf{b}}(s,a), N_h(s,a)\big\}}, H \right\} \tag{105}$$

  for some universal constant $c_{\mathsf{b}}>0$. Here, $\mathsf{Var}_{\widehat{P}_{h,s,a}}\big(\widehat{V}_{h+1}\big)$ corresponds to the variance of $\widehat{V}_{h+1}$ w.r.t. the distribution $\widehat{P}_{h,s,a}$.

- In the reward-free setting, the penalty term is taken as

$$b_h(s,a) = \min\left\{\sqrt{\frac{c_{\mathsf{b}} S \log\frac{HSA}{\delta}}{\min\left\{\widehat{N}_h^{\mathsf{b}}(s,a), N_h(s,a)\right\}}\mathsf{Var}_{\widehat{P}_{h,s,a}}\left(\widehat{V}_{h+1}\right)} + c_{\mathsf{b}} SH\frac{\log\frac{HSA}{\delta}}{\min\left\{\widehat{N}_h^{\mathsf{b}}(s,a), N_h(s,a)\right\}}, H\right\}. \tag{106}$$

With such penalty terms in place, we are ready to present the whole model-based offline RL algorithm in Algorithm 4.

**Algorithm 4:** Pessimistic model-based offline RL.

**input:** a dataset $\mathcal{D} = \left\{(s_h^{n,\mathsf{b}}, a_h^{n,\mathsf{b}}, s_{h+1}^{n,\mathsf{b}})\right\}_{1\le n\le K, 1\le h<H}$; reward function $r$.
**initialization:** $\widehat{V}_{H+1} = 0$.
**subsampling:** compute the lower bound of the number of sample transitions $\widehat{N}^{\mathsf{b}}$ according to (22);
subsample $\mathcal{D}$ to obtain $\mathcal{D}^{\mathsf{trim}}$, such that for each $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, $\mathcal{D}^{\mathsf{trim}}$ contains $\min\left\{\widehat{N}_h^{\mathsf{b}}(s,a), N_h(s,a)\right\}$ sample transitions randomly drawn from $\mathcal{D}$.
**for** $h = H, \cdots, 1$ **do**
  compute the empirical transition kernel $\widehat{P}_h$ according to (104).
  **for** $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    compute the penalty term $b_h(s,a)$ according to (105) for the reward-agnostic case or (106) for the reward-free case.
    set $\widehat{Q}_h(s,a) = \max\left\{r_h(s,a) + \widehat{P}_{h,s,a}\widehat{V}_{h+1} - b_h(s,a), 0\right\}$.
  **for** $s \in \mathcal{S}$ **do**
    set $\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a)$ and $\widehat{\pi}_h(s) = \arg\max_a \widehat{Q}_h(s,a)$.
**output:** $\widehat{\pi} = \{\widehat{\pi}_h\}_{1\le h\le H}$.

# B Proof of Lemma 1

If $g\big(\pi^{(t)}, \widehat{d}, \mu_{\mathsf{b}}^{(t)}\big) \ge 2HSA$, then the learning rate necessarily obeys

$$\frac{1}{2HSA-1} \le \alpha_t < \frac{1}{HSA}. \tag{107}$$

This combined with the update rule (31) allows one to lower bound the progress made in each iteration:

$$\begin{aligned}
&\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \left\{\log\left[\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu_{\mathsf{b}}^{(t+1)}}\left[\widehat{d}_h^{\pi}(s,a)\right]\right] - \log\left[\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu_{\mathsf{b}}^{(t)}}\left[\widehat{d}_h^{\pi}(s,a)\right]\right]\right\} \\
&= \sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \log\left[\frac{\frac{1}{KH} + \mathbb{E}_{\pi\sim(1-\alpha_t)\mu_{\mathsf{b}}^{(t)}(\pi)+\alpha_t\,\mathbb{1}(\pi^{(t)})}\left[\widehat{d}_h^{\pi}(s,a)\right]}{\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu_{\mathsf{b}}^{(t+1)}}\left[\widehat{d}_h^{\pi}(s,a)\right]}\right] \\
&= \sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \log\left[1 + \alpha_t\left(\frac{\frac{1}{KH} + \widehat{d}_h^{\pi^{(t)}}(s,a)}{\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu_{\mathsf{b}}^{(t+1)}}\left[\widehat{d}_h^{\pi}(s,a)\right]} - 1\right)\right] \\
&\overset{\text{(i)}}{\ge} \log\left[1 + \alpha_t\Big(g\big(\pi^{(t)}, \widehat{d}, \mu_{\mathsf{b}}^{(t)}\big) - 1\Big)\right] + (HSA-1)\log(1-\alpha_t) \\
&\overset{\text{(ii)}}{\ge} \log\left[1 + \alpha_t\big(2HSA-1\big)\right] + (HSA-1)\log(1-\alpha_t) \\
&\overset{\text{(iii)}}{\ge} \min\left\{\log 2 + (HSA-1)\log\left(1 - \frac{1}{2HSA-1}\right), \log\left(3 - \frac{1}{HSA}\right) + (HSA-1)\log\left(1 - \frac{1}{HSA}\right)\right\}
\end{aligned}$$

$$\overset{\text{(iv)}}{\geq} 0.09, \tag{108}$$

where (ii) arises since $g\big(\pi^{(t)}, \widehat{d}, \mu_{\mathsf{b}}^{(t)}\big) \geq 2HSA$, (iii) makes use of (107), and (iv) is derived by numerically calculating the function $\min\big\{\log 2 + (x-1)\log\big(1-\frac{1}{2x-1}\big), \log\big(3-\frac{1}{x}\big) + (x-1)\log\big(1-\frac{1}{x}\big)\big\}$ for $x > 1$. To see why (i) holds, we make note of the following fact that holds for any $x_1, x_2 \geq 0$ and any $\alpha \in (0,1)$:

$$\begin{aligned}\log\big[1+\alpha(x_1-1)\big] + \log\big[1+\alpha(x_2-1)\big] &= \log\big[1+\alpha(x_1+x_2-2)+\alpha^2(x_1-1)(x_2-1)\big] \\ &\geq \log\big[1+\alpha(x_1+x_2-2)-\alpha^2(x_1+x_2-1)\big] \\ &= \log\big[1+\alpha(x_1+x_2-1)\big] + \log\big[1-\alpha\big],\end{aligned}$$

which in turn implies that

$$\sum_{i=1}^{n}\log\big[1+\alpha(x_i-1)\big] \geq \log\bigg[1+\alpha\Big(\sum_{i=1}^{n}x_i-1\Big)\bigg] + (n-1)\log\big[1-\alpha\big].$$

In addition, it is straightforward to check that the objective function satisfies

$$-HSA\log(KH) \leq f(\mu) = \sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}\log\bigg[\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu}\big[\widehat{d}_h^{\pi}(s,a)\big]\bigg] \leq HSA\log 2$$

for any $\mu \in \Delta(\Pi)$. Taking this collectively with (108), we immediately see that the subroutine terminates within $O\big(HSA\log(KH)\big)$ iterations.

# References


Agarwal, A., Kakade, S., and Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. *Conference on Learning Theory*, pages 67–83.

Auer, P. and Ortner, R. (2006). Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. (2019). Provably efficient $q$-learning with low switching cost. In *Advances in Neural Information Processing Systems*, pages 8002–8011.

Bertsekas, D. P. (2017). *Dynamic programming and optimal control (4th edition)*. Athena Scientific.

Brafman, R. I. and Tennenholtz, M. (2002). R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231.

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

Chen, X., Hu, J., Yang, L. F., and Wang, L. (2021). Near-optimal reward-free exploration for linear mixture MDPs with plug-in solver. *arXiv preprint arXiv:2110.03244*.

Cui, Q. and Du, S. S. (2022a). Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. *arXiv preprint arXiv:2206.00159*.

Cui, Q. and Du, S. S. (2022b). When is offline two-player zero-sum Markov game solvable? *arXiv preprint arXiv:2201.03522*.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021). Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598.

Dong, K., Wang, Y., Chen, X., and Wang, L. (2019). Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311*.

Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W., and Wang, R. (2021). Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*.

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110.

Gelfand, I. M., Silverman, R. A., et al. (2000). *Calculus of variations*. Courier Corporation.

Huang, R., Yang, J., and Liang, Y. (2022). Safe exploration incurs nearly no additional sample complexity for reward-free RL. *arXiv preprint arXiv:2206.14057*.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. (2020a). Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020b). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143.

Jin, Y., Ren, Z., Yang, Z., and Wang, Z. (2022). Policy learning "without" overlap: Pessimism and generalized empirical Bernstein's inequality. *arXiv preprint arXiv:2212.09900*.

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096.

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. (2021). Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pages 865–891. PMLR.

Kearns, M. and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Li, G., Chen, Y., Chi, Y., Gu, Y., and Wei, Y. (2021). Sample-efficient reinforcement learning is feasible for linearly realizable MDPs with limited revisiting. *Advances in Neural Information Processing Systems*, 34:16671–16685.

Li, G., Shi, L., Chen, Y., and Chi, Y. (2022a). Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *accepted to Information and Inference*.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2022b). Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*.

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2023). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *accepted to Operations Research*.

Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. (2021a). Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pages 7599–7608.

Ménard, P., Domingues, O. D., Shang, X., and Valko, M. (2021b). UCB momentum Q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pages 7609–7618.

Miryoosefi, S. and Jin, C. (2022). A simple reward-free approach to constrained reinforcement learning. In *International Conference on Machine Learning*, pages 15666–15698.

Qiao, D. and Wang, Y.-X. (2022). Near-optimal deployment efficiency in reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2210.00701*.

Qiao, D., Yin, M., Min, M., and Wang, Y.-X. (2022). Sample-efficient reinforcement learning with $\log\log(t)$ switching cost. *arXiv preprint arXiv:2202.06385*.

Qiu, S., Ye, J., Wang, Z., and Yang, Z. (2021). On reward-free RL with kernel and neural function approximations: Single-agent mdp and markov game. In *International Conference on Machine Learning*, pages 8737–8747.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022). Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. *International Conference on Machine Learning*.

Simchowitz, M. and Jamieson, K. G. (2019). Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. (2022). Reward-free RL is no harder than reward-aware RL in linear Markov decision processes. In *International Conference on Machine Learning*, pages 22430–22456.

Wang, R., Du, S. S., Yang, L., and Salakhutdinov, R. R. (2020). On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 33:17816–17826.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407.

Xu, T. and Liang, Y. (2022). Provably efficient offline reinforcement learning with trajectory-wise reward. *arXiv preprint arXiv:2206.06426*.

Yan, Y., Li, G., Chen, Y., and Fan, J. (2022a). The efficacy of pessimism in asynchronous Q-learning. *arXiv preprint arXiv:2203.07368*.

Yan, Y., Li, G., Chen, Y., and Fan, J. (2022b). Model-based reinforcement learning is minimax-optimal for offline zero-sum Markov games. *arXiv preprint arXiv:2206.04044*.

Yin, M., Bai, Y., and Wang, Y.-X. (2021). Near-optimal offline reinforcement learning via double variance reduction. *arXiv preprint arXiv:2102.01748*.

Yin, M. and Wang, Y.-X. (2021a). Optimal uniform OPE and model-based offline reinforcement learning in time-homogeneous, reward-free and task-agnostic settings. *Advances in neural information processing systems*, 34:12890–12903.

Yin, M. and Wang, Y.-X. (2021b). Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34:4065–4078.

Zanette, A. and Brunskill, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312.

Zanette, A., Lazaric, A., Kochenderfer, M. J., and Brunskill, E. (2020). Provably efficient reward-agnostic navigation with linear value iteration. *Advances in Neural Information Processing Systems*, 33:11756–11766.

Zhang, W., Zhou, D., and Gu, Q. (2021a). Reward-free model-based reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34:1582–1593.

Zhang, X., Ma, Y., and Singla, A. (2020a). Task-agnostic exploration in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:11734–11743.

Zhang, Z., Du, S., and Ji, X. (2021b). Near optimal reward-free reinforcement learning. In *International Conference on Machine Learning*, pages 12402–12412. PMLR.

Zhang, Z., Zhou, Y., and Ji, X. (2020b). Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33.