Achieving $\tilde{O}(1/\varepsilon)$ Sample Complexity for Constrained Markov Decision Process

Jiashuo Jiang
Department of Industrial Engineering & Decision Analytics
Hong Kong University of Science and Technology
Hong Kong, China
jsjiang@ust.hk

Yinyu Ye
Department of Management Science & Engineering
Institue of Computational Mathematics and Engineering
Stanford University
California, US
yyye@stanford.edu

Abstract

We consider the reinforcement learning problem for the constrained Markov decision process (CMDP), which plays a central role in satisfying safety or resource constraints in sequential learning and decision-making. In this problem, we are given finite resources and a MDP with unknown transition probabilities. At each stage, we take an action, collecting a reward and consuming some resources, all assumed to be unknown and need to be learned over time. In this work, we take the first step towards deriving optimal problem-dependent guarantees for the CMDP problems. We derive a logarithmic regret bound, which translates into a $O(\frac{1}{\Delta\cdot\varepsilon}\cdot\log^2(1/\varepsilon))$ sample complexity bound, with Δ being a problem-dependent parameter, yet independent of ε. Our sample complexity bound improves upon the state-of-art $O(1/\varepsilon^2)$ sample complexity for CMDP problems established in the previous literature, in terms of the dependency on ε. To achieve this advance, we develop a new framework for analyzing CMDP problems. To be specific, our algorithm operates in the primal space and we resolve the primal LP for the CMDP problem at each period in an online manner, with *adaptive* remaining resource capacities. The key elements of our algorithm are: i) a characterization of the instance hardness via LP basis, ii) an eliminating procedure that identifies one optimal basis of the primal LP, and; iii) a resolving procedure that is adaptive to the remaining resources and sticks to the characterized optimal basis.

1 Introduction

Reinforcement learning (RL) is pivotal in the realm of dynamic decision-making under uncertainty, where the objective is to maximize total reward through ongoing interaction with and learning from an enigmatic environment. Markov Decision Processes (MDPs) are a prevalent framework for encapsulating environmental dynamics. MDPs have been instrumental in various domains, such as video gaming [52], robotics [41], recommender systems [57], inventory control [11], and beyond. Yet, they fall short in accommodating additional constraints that may influence the formulation of the optimal policy and the decision-maker's engagement with the uncertain environment. Often, in MDP applications, there are stringent constraints on utilities or costs, emanating from areas like safe

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

autonomous driving [22], robotics [56], revenue management [33] and financial management [62]. These constraints might also symbolize limitations on resources within resource allocation applications. Constrained MDPs (CMDPs), as introduced in [4], enhance MDPs to factor in constraints affecting long-term policy results. In CMDPs, the decision-maker aims to optimize cumulative rewards while adhering to these constraints. Our paper focuses on CMDPs, and we aim to develop efficient algorithmic solutions.

The significance of RL in CMDP contexts has garnered substantial attention in recent years. A variety of methods for tackling CMDPs have been developed, including the primal-dual technique [21, 49, 12, 69, 18, 45, 25, 26], which leverages the Lagrangian dual of CMDPs and implements an online learning strategy for the iterative update of dual variables. Other methods encompass constrained optimization [1, 61], the Lyapunov technique [14], among others. Previous research has established *minimax* bounds for CMDPs, representing the optimal regret that can be achieved for the most challenging problem within a specific class of problems. Nonetheless, these minimax regret bounds and worst-case scenarios can be overly conservative, leading to a disconnect between theoretical guarantees and practical performance for specific problem instances. A more tailored approach is warranted—one that ensures great performance on every single problem instance and furnishes problem-dependent guarantees. Our research takes the first step towards deriving optimal problem-dependent guarantees for CMDP problems.

1.1 Preliminaries

We consider a CMDP problem with a finite set of states $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$ and a finite set of actions $\mathcal{A} = \{1, 2, \ldots, |\mathcal{A}|\}$. We denote by $\gamma \in (0, 1)$ a discount factor. We also denote by $P : \mathcal{S} \times \mathcal{A} \to \mathcal{D}(\mathcal{S})$ the probability transition kernel of the CMDP, where $\mathcal{D}(\mathcal{S})$ denotes a probability measure over the state space $\mathcal{S}$. Then, $P(s'|s, a)$ denotes the probability of transiting from state s to state s' when the action a is executed. The initial distribution over the states of the CMDP is denoted by μ_1.

There is a *stochaastic* reward function $r : \mathcal{S} \times \mathcal{A} \to \mathcal{D}[0, 1]$ and K *stochastic* cost functions $c_k : \mathcal{S} \times \mathcal{A} \to \mathcal{D}[0, 1]$ for each $k \in [K]$. We also denote by $\hat{r}(s, a) = \mathbb{E}[r(s, a)]$ for each (s, a) and $\hat{c}(s, a) = \mathbb{E}[c_k(s, a)]$ for each (s, a). For any Markovian policy π, where the action of π depends only on the current state and the action of π is allowed to be randomized, we denote by $V_r(\pi, \mu_1)$ the infinite horizon discounted reward of the policy π, with the formulation of $V_r(\pi, \mu_1)$ given below:

$$V_r(\pi, \mu_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r(s_t, a_t) \mid \mu_1\right], \tag{1}$$

where (s_t, a_t) is generated according to the policy π and the transition kernel P with the initial state distribution μ_1. For each $k \in [K]$, the infinite horizon discounted cost of the policy π is denoted by $V_k(\pi, \mu_1)$, and the following constraint needs to be satisfied by the policy π,

$$V_k(\pi, \mu_1) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot c_k(s_t, a_t) \mid \mu_1\right] \leq \alpha_k, \quad \forall k \in [K]. \tag{2}$$

To solve the CMDP problem, we aim to find an optimal Markovian policy, denoted by π^*, that maximizes the reward in (1) while satisfying the cost constraint (2) for each $k \in [K]$, with $\alpha_k \in \left[0, \frac{1}{1-\gamma}\right]$ being a pre-specified value for each $k \in [K]$. Importantly, we assume that the reward function r, the cost functions $\{c_k\}_{k=1}^K$, and the transition kernel P, are all **unknown** to the decision maker. Our goal is to obtain a policy π that approximates the optimal policy π^* with as few samples as possible. We now describe the sampling procedure and present the performance measure of our policy. We assume the existence of a stylized *generative model* $\mathcal{M}$, as studied in [39, 38, 44]. The model $\mathcal{M}$ satisfies the following condition.

Assumption 1.1 *For each state and action pair (s, a), we can query the model $\mathcal{M}$ to obtain an observation of $r(s, a)$, $c_k(s, a)$ for each $k \in [K]$, and the new state $s' \in \mathcal{S}$, where the transition from s to s' follows the probability kernel $P(s'|s, a)$ independently.*

Note that in reinforcement learning for CMDP problems, querying the generative model $\mathcal{M}$ can be costly. Therefore, it is desirable to query the model $\mathcal{M}$ as less as possible, while guaranteeing the

near optimality of the approximate policy. To this end, we use *sample complexity* as the measure of the performance of our policy. For any ε, we aim to find an ε-accurate policy π such that

$$V_r(\pi^*, \mu_1) - V_r(\pi, \mu_1) \leq \varepsilon \text{ and } V_k(\pi, \mu_1) - \alpha_k \leq \varepsilon, \forall k \in [K], \tag{3}$$

with as few samples as possible.

1.2 Our Main Results and Contributions

The main result of our research is the introduction of a novel algorithm that promises a $O(\frac{1}{\Delta \cdot \varepsilon} \cdot \log^2(1/\varepsilon))$ sample complexity bound, where Δ is a positive constant that characterizes the gap between the optimal policy and the sub-optimal ones. Note that the state-of-the-art sample complexity bound under the worst-case scenarios is $O(1/\varepsilon^2)$, which has been established in a series of work [68, 37, 21]. Though the $O(1/\varepsilon)$ or better iteration complexity has been achieved in [45, 50, 74, 26], it comes with a sample complexity bound no better than $O(1/\varepsilon^2)$. Our algorithm enjoys a sample complexity bound that has a better dependency in terms of ε. To achieve this improved result, we develop several new elements listed below.

Contribution 1: we develop new characterizations of the problem instance hardness for CMDP problems. Note that a key component for achieving instance-dependent bounds is to characterize the "hardness" of the underlying problem instance. That is, we need to identify a positive gap to separate the optimal policy from the sub-optimal policies for a particular problem instance. The importance of identifying such a gap has been demonstrated in instance-optimal learning for multi-arm-bandit problems (e.g. [42]) and reinforcement learning problems (e.g. [6]), among others. For CMDP, identifying such a gap is non-trivial because the optimal policies for CMDP are randomized policies [4]. Then, the policies can be represented by distributions over the action set and the sub-optimal policies can be arbitrarily close to the optimal policy as long as the corresponding sub-optimal distributions converge to the optimal one. To tackle this problem, we show that the feasible region for the policies can be represented as a polytope and we only need to focus on the corner points of this polytope to find an optimal policy. Therefore, the hardness can simply be characterized as the distance between the optimal corner point and the sub-optimal corner point, as illustrated in detail in Section 2.1. This is the first characterization of problem instance hardness for CMDP problems.

Contribution 2: we devise a new algorithmic framework to analyze CMDP problems, inspired by the online packing/linear programming (LP) literature [3, 40, 47, 46]. Specifically, we utilize a linear programming reformulation of the CMDP problem, where policies are delineated via *occupancy measures* [4]. The optimal policy emerges from the LP's solution; however, the indeterminate model parameters mean the LP cannot be solved directly but must be approached online as we obtain more samples from the generative model. Each generative model query leads to solving an empirical LP with accrued samples, and our final policy is derived from averaging these solutions—an approach akin to the methodology in online LP. A critical feature of our algorithm is the adaptiveness of the LP constraints' right-hand side to the input samples, a technique demonstrated to achieve logarithmic regret in online LP literature, which we now apply to CMDP problems.

Contribution 3: we extend our contributions to the online LP literature. Note that after adopting the LP reformulation, the corner points of the feasible region for policies can be represented by the basis of the LP. Separating the optimal policy from the others is equivalent to identifying one optimal basis of the LP. We utilize an approach that lexicographically restricts one variable to zero and tests whether the LP value has changed. We show that this approach systematically pinpoints a particular optimal LP basis with a high probability. Then, we develop a resolving procedure that capitalizes on the structure of the identified optimal basis, which involves only the non-zero basic variables and the active constraints. This is a new approach of deriving problem-dependent bound for online LP.

Other related literature. Problem-dependent guarantees have been considered extensively in the RL literature, where a series of work [72, 59, 54, 23, 17, 66, 65, 71, 20] establishes the $\log(N)$ regret bound or the $O(\kappa \cdot \varepsilon^{-1})$ sample complexity, with κ being a problem-dependent constant. Our approach can also be directly applied to the RL problems. Moreover, our approach can handle long-term constraints and can deal with multi-objective (safe) RL problems.

our work presents a new algorithm for safe RL problems. We adopt an occupancy measure representation of the optimal and obtain an LP to work with, which is similar to the previous work. However, our algorithm resolves an LP and operates in the primal space, which is fundamentally different from the previous work that adopts a primal-dual update (e.g. [60], [19], [73], [8], [53]). There is also

work developing primal-based algorithms, for example ([51], [13], [15], [16], [70]). Our algorithm is completely different from the previous work and we obtain new results. The result is that we are able to obtain an instance-dependent $\tilde{O}(1/\epsilon)$ sample complexity the first time in the literature, which improves upon the $O(1/\epsilon^2)$ worst-case sample complexity established in the previous work. Though the constrained optimization approach and the Lyapunov approach have also been developed for CMDP problems, they do not enjoy a theoretical guarantee. In comparison to the literature, we develop a new primal-based algorithm and achieve the first instance-dependent sample complexity for CMDP problems.

Our new primal-based algorithm is motivated from the literature of online linear programming [3, 40, 47] and bandits with knapsack problems (e.g. [46], [48]). In these problems, the optimal policy can be written as an LP and we need to develop an online policy to solve this LP sequentially. Note that a prevalent strategy is to resolve the LP adaptive to the remaining resources, which has been developed in a long line of research on various applications, for example [31], [32], [5], [36], [64], [35] and [10]. We make the innovation of resolving the LP while sticking to the identified optimal basis, which distinguishes our algorithm from the previous ones in online LP. Note that the online LP techniques has also been extended to handle non-stationarity, for example in [9], [34]. It is an interesting future topic to explore whether our algorithm can be extended to non-stationary environment.

2 LP Reformulation

The infinite horizon discounted setting described in Section 1.1 admits a linear programming reformulation. To be specific, due to the existence of the constraints, the optimal policy of a CMDP can be randomized policies (e.g. [4]), where it is optimal to take a stochastic action given the current state. Therefore, it is convenient to represent a policy through the *occupation measure*, which gives us the desired linear programming reformulations of the CMDP problems.

For the infinite horizon discounted problem, the occupancy measure is defined as $q_\pi(s,a)$ for any state s, action a, and policy π. Note that $q_\pi(s,a)$ represents the total expected discounted time spent on the state-action pair (s,a), under policy π, multiplied by $1-\gamma$. Then, following [4], the optimal policy (and the optimal occupancy measure) can be obtained from the following linear programming.

$$
\begin{aligned}
V^{\text{Infi}} = \quad & \max \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \hat{r}(s,a)\cdot q(s,a) && && \text{(4a)}\\
& \text{s.t.} \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \hat{c}_k(s,a)\cdot q(s,a) \le \alpha_k && \forall k\in[K] && \text{(4b)}\\
& \sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} q(s',a)\cdot(\delta_{s,s'} - \gamma\cdot P(s|s',a)) = (1-\gamma)\cdot\mu_1(s) && \forall s\in\mathcal{S} && \text{(4c)}\\
& q(s,a)\ge 0 && \forall s\in\mathcal{S}, a\in\mathcal{A}, && \text{(4d)}
\end{aligned}
$$

where $\delta_{s,s'} = \mathbb{1}_{s=s'}$, and $\mu_1(s)$ denotes the probability for the first state to be realized as $s\in\mathcal{S}$ following the initial distribution μ_1. Note that for an optimal solution $\{q^*(s,a)\}_{\forall s\in\mathcal{S},\forall a\in\mathcal{A}}$ to (4), the corresponding optimal policy π^* will be

$$
P(\pi^*(s)=a) = \begin{cases} \dfrac{q^*(s,a)}{\sum_{a'\in\mathcal{A}} q^*(s,a')}, & \text{if } \sum\limits_{a'\in\mathcal{A}} q^*(s,a') > 0 \\ 1/|\mathcal{A}|, & \text{if } \sum\limits_{a'\in\mathcal{A}} q^*(s,a') = 0, \end{cases} \tag{5}
$$

where $\pi^*(s)$ denotes the probability for the policy π^* to take the action $a\in\mathcal{A}$ given the state $s\in\mathcal{S}$. In fact, when $\sum_{a'\in\mathcal{A}} q^*(s,a') = 0$, we can take an arbitrary action. In what follows, we rely on the linear programming formulation (4) to derive our results.

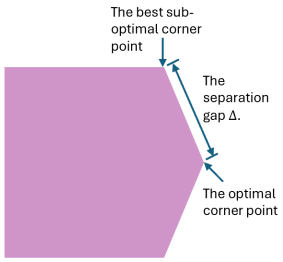

Figure 1: A graph illustration of the hardness characterization via LP basis, where the shaded area denotes the feasible region for the policies.

2.1 Characterization of Instance Hardness

We now rewrite the LP (4) into the following standard formulation to proceed with our illustration.

$$\begin{aligned} V = \max \quad & \hat{\boldsymbol{r}}^\top \boldsymbol{q} \\ \text{s.t.} \quad & C\boldsymbol{q} \leq \boldsymbol{\alpha} \\ & B\boldsymbol{q} = \boldsymbol{\mu} \\ & \boldsymbol{q} \geq 0. \end{aligned} \tag{6}$$

Note that one crucial step for achieving problem-dependent bounds is to characterize the hardness of the underlying problem instance and define a gap that separates the optimal policies from the others. For multi-arm-bandit problem, the characterization of the hardness can be the gap between the optimal arm and the best sub-optimal arm (e.g. [42]). For reinforcement learning problem, the characterization of the hardness can also be the gap between the optimal policy and the best sub-optimal policy (e.g. [6]). As long as this separation gap is a positive constant, denoted by Δ, separating the optimal policy from the others with a probability at least $1-\epsilon$ would require samples at most $\frac{1}{\Delta} \cdot \log(1/\epsilon)$, which finally implies the instance-optimal sample complexity bound.

For CMDP problem, characterizing the hardness of the problem instance can be hard. Based on LP (6), we know that a feasible policy corresponds to a feasible solution and the sub-optimal solution can be arbitrarily close to the optimal solution since the feasible set is "continuous". Therefore, there is no direct way to identify a positive gap between the optimal policies and the sub-optimal ones. However, from standard LP theory, we know that one **corner point** of the feasible region must be one optimal solution. Therefore, we can simply focus on the corner points when solving LP (6) and we define the gap as *the distance between the optimal corner point and the sub-optimal corner point*, as illustrated in Figure 1. As we will show later, the LP reformulation (6) and such a characterization of hardness via corner points will inspire our entire approach.

Since the problem hardness is characterized via corner points, it is essential to provide further characterization of the corner points. Note that in LP theory, the corner point is called *basic solution* and can be represented by *LP basis*, which involves the set of basic variables that are allowed to be non-zero, and the set of active constraints that are binding under the corresponding basic solution. Our next lemma follows from standard LP theory, where the proof is provided in the appendix for completeness.

Lemma 2.1 *Denote by b the number of rows in the matrix B. Then, there exists a subset $\mathcal{J}^* \subset [K]$ with $K' = |\mathcal{J}^*|$ and an optimal solution $\boldsymbol{q}^*$ to LP (6) such that there are $b + K'$ variables in $\boldsymbol{q}^*$ are non-zero. Moreover, we denote by $\mathcal{I}^*$ the index set of the non-zero element in $\boldsymbol{q}^*$. Then, the optimal solution $\boldsymbol{q}^*$ can be uniquely determined as the solution to the linear system*

$$C(\mathcal{J}^*, \mathcal{I}^*)\boldsymbol{q}_{\mathcal{I}^*} = \boldsymbol{\alpha}_{\mathcal{J}^*}, \tag{7a}$$

$$B(:, \mathcal{I}^*)\boldsymbol{q}_{\mathcal{I}^*} = \boldsymbol{\mu}, \tag{7b}$$

$$\boldsymbol{q}_{\mathcal{I}^{*c}} = 0. \tag{7c}$$

with $\mathcal{I}^{*c}$ being the complementary set of the index set $\mathcal{I}^*$.

Remark. Note that the minimax lower bound $\Omega(1/\epsilon^2)$ has been established in previous work [7, 63]. However, this does not contradict with our $\tilde{O}(1/\epsilon)$ sample complexity after we introduce the instance

hardness measure Δ. To be specific, for a problem instance I, we can denote by $S(I,\epsilon)$ the number of samples needed to construct an ϵ-optimal policy. Then the worst-case lower bound implies that $\max_I S(I,\epsilon) = \Theta(1/\epsilon^2)$. However, if we do not consider the worst-case guarantee, i.e., if we do not maximize over the problem instance I, then we can characterize an instance-dependent constant $\Delta(I)$ (independent of ϵ) such that $S(I,\epsilon) = \Delta(I)/\epsilon \cdot \text{polylog}(1/\epsilon)$. When the problem instance is favorable such that the constant $\Delta(I)$ is smaller than $1/\epsilon$, our bound strictly improves upon the worst-case bound.

Overview of our approach: in the first part, we aim to identify the optimal basis $\mathcal{I}$ and $\mathcal{J}$. In this way, we identify the optimal corner point to look at. The detailed procedure is described in Section 3. In the second part, we learn the optimal solution given the optimal basis we have identified, which finally gives us a near-optimal policy with the desired sample complexity bound. The detailed procedure is described in Section 4.

3 Construct Estimates and Identify Optimal Basis

We describe how to construct estimates for LP (6). To this end, for a round N_0, we denote by $\mathcal{F}_{N_0}$ the filtration of all the information collected up to round N_0. Then, we denote by $\bar{C}_{N_0}$ (resp. $\bar{B}_{N_0}$) an estimate of the matrix C (resp. B), constructed using the information in the set $\mathcal{F}_{N_0}$. We also denote by $\bar{\boldsymbol{r}}_{N_0}$ and estimate of $\hat{\boldsymbol{r}}$ constructed from the information in the set $\mathcal{F}_{N_0}$. To be specific, we define

$$\bar{r}_{N_0}(s,a) = \frac{\sum_{n=1}^{N_0} r^n(s,a)}{N_0}, \bar{c}_{k,N_0}(s,a) = \frac{\sum_{n=1}^{N_0} c_k^n(s,a)}{N_0}, \text{ and } \bar{P}_{N_0}(s'|s,a) = \frac{\sum_{n=1}^{N_0} \mathbb{1}_{s^n(s,a)=s'}}{N_0}, \tag{8}$$

where $r^n(s,a)$ denotes the n-th observation of the reward, and $c_k^n(s,a)$ denotes the n-th observation of the k-th cost, and $s^n(s,a)$ denotes the n-th observation of the state transition for the state-action pair (s,a), for $n \in [N_0]$. Then, similar to [21], we can use the following LP to obtain an estimate of V (6).

$$\begin{aligned} \bar{V}_{N_0} = \max \quad & (\bar{\boldsymbol{r}}_{N_0})^\top \boldsymbol{q} \\ \text{s.t.} \quad & \bar{C}_{N_0}\boldsymbol{q} \le \boldsymbol{\alpha} + \lambda_{N_0} \\ & \bar{B}_{N_0}\boldsymbol{q} \le \boldsymbol{\mu} + \lambda_{N_0} \\ & \bar{B}_{N_0}\boldsymbol{q} \ge \boldsymbol{\mu} - \lambda_{N_0} \\ & \boldsymbol{q} \ge 0, \end{aligned} \tag{9}$$

with λ_{N_0} being a parameter that we specify later. To bound the estimation gap between $\bar{V}_{N_0}$ and V, it is useful to bound the optimal dual solution to (6). To this end, we adopt the approach in [29, 55] that utilizes Slater's condition, which is imposed as an assumption below.

Assumption 3.1 *There exists a policy $\bar{\pi}$ such that all the resource constraints are satisfied strictly. In other words, there exists an occupancy measure $\bar{\boldsymbol{q}}$ such that $B\bar{\boldsymbol{q}} = \boldsymbol{\mu}$ and $C\bar{\boldsymbol{q}} < \boldsymbol{\alpha}$. In fact, for each state $s \in \mathcal{S}$, there exists a null action that consumes no resource.*

The Slater point $\bar{\boldsymbol{q}}$ can be set as the policy that takes the null action given each state. The estimation error will be related to the gap between the Slater point $\bar{\boldsymbol{q}}$ and the optimal point $\boldsymbol{q}^*$. We then define the lower gap as

$$\text{Gap}_1(N_0,\varepsilon) \ge V - \bar{V}_{N_0} \tag{10}$$

and the upper gap as

$$\text{Gap}_2(N_0,\varepsilon) \ge \bar{V}_{N_0} - V \tag{11}$$

with both inequalities (10) and (11) hold with probability at least $1-\varepsilon$.

3.1 Bound the Estimation Gap

We denote by $\text{Rad}(N_0,\varepsilon) = \sqrt{\frac{\log(2/\varepsilon)}{2N_0}}$. Following the standard Hoeffding's inequality, we know that $|\bar{r}_{N_0}(s,a) - \hat{r}(s,a)|$, $|\bar{c}_{k,N_0}(s,a) - \hat{c}_k(s,a)|$, and $|\bar{P}_{N_0}(s'|s,a) - P(s'|s,a)|$ are all upper bounded by $\text{Rad}(N_0,\varepsilon)$ with probability at least $1-\varepsilon$. We can simply set

$$\text{Gap}_1(N_0,\varepsilon) = \text{Rad}(N_0,\varepsilon) \tag{12}$$

and

$$\mathrm{Gap}_2(N_0,\varepsilon)=\frac{2\mathrm{Rad}(N_0,\varepsilon)}{\min_{k\in[K]}\{\alpha_k\}}\cdot\left(1+\frac{|\mathcal{S}|}{1-\gamma}\right)+\frac{\mathrm{Rad}^2(N_0,\varepsilon)}{\min_{k\in[K]}\{\alpha_k\}}\cdot\left(|\mathcal{S}|+\frac{|\mathcal{S}|^2}{1-\gamma}\right). \tag{13}$$

We have the following result, where the proof is relegated to appendix.

Lemma 3.2 *As long as* $\lambda_{N_0}=\mathrm{Rad}(N_0,\varepsilon)$*, the following inequality*

$$V\le\bar{V}_{N_0}+\mathrm{Gap}_1(N_0,\varepsilon)\le V+\mathrm{Gap}_1(N_0,\varepsilon)+\mathrm{Gap}_2(N_0,\varepsilon). \tag{14}$$

holds with probability at least $1-(K|\mathcal{S}||\mathcal{A}|-|\mathcal{S}|^2|\mathcal{A}|)\cdot\varepsilon$*, where* $\mathrm{Gap}_1(N_0,\varepsilon)$ *is defined in* (12) *and* $\mathrm{Gap}_2(N_0,\varepsilon)$ *is defined in* (13)*.*

3.2 Characterize One Optimal Basis

We now describe how to identify one optimal basis of the LP (6) as required in Lemma 2.1, by sequentially discarding the sub-optimal actions and the redundant constraints. The formal algorithm to identify such non-zero elements and the constraints is given in Algorithm 1.

Algorithm 1 Algorithm for identifying one optimal basis

1: **Input:** the historical sample set $\mathcal{F}_{N_0}$ that contains N_0 samples for each $(s,a)\in\mathcal{S}\times\mathcal{A}$.
2: Compute the value of $\bar{V}_{N_0}$ as in (9).
3: Initialize $\mathcal{I}$ to be the whole index set that contains every column index of matrix B in (6) and $\mathcal{J}=[K]$.
4: **for** $i\in\mathcal{I}$ **do**
5: Let $\mathcal{I}'=\mathcal{I}\backslash\{i\}$.
6: Compute the value of $\bar{V}_{\mathcal{I}',N_0}$ as in (15).
7: If $|\bar{V}_{\mathcal{I}',N_0}-\bar{V}_{N_0}|\le 2\mathrm{Gap}_1(N_0,\varepsilon)+2\mathrm{Gap}_2(N_0,\varepsilon)$, then we set $\mathcal{I}=\mathcal{I}'$.
8: **end for**
9: **for** $k=1,\ldots,K$ **do**
10: Let $\mathcal{J}'=\mathcal{J}\backslash\{q\}$.
11: Compute the value of $\mathrm{Dual}_{\mathcal{J}',\mathcal{I},N_0}$ as in (18).
12: If $|\bar{V}_{N_0}-\mathrm{Dual}_{\mathcal{J}',\mathcal{I},N_0}|\le 2\mathrm{Gap}_1(N_0,\varepsilon)+2\mathrm{Gap}_2(N_0,\varepsilon)$, then we set $\mathcal{J}=\mathcal{J}'$.
13: **end for**
14: **Output**: the set of indexes $\mathcal{I}$ and $\mathcal{J}$.

We now explain the intuition why Algorithm 1 works. Denote by $\boldsymbol{q}^*$ an optimal solution to LP (6). We describe how to identify the non-zero elements in $\boldsymbol{q}^*$ and how to identify the constraints such that the values of the non-zero elements of $\boldsymbol{q}^*$ can be uniquely determined by the corresponding linear equation. For each i-th element of $\boldsymbol{q}^*$, we compare the value of V (6) against V with an additional constraint that $q_i=0$. If the two values are different, we identify a non-zero element. To this end, for an index set $\mathcal{I}$, we define an LP, as well as its estimate, as follows.

$$\begin{array}{rlll} V_{\mathcal{I}}=\max & \hat{\boldsymbol{r}}^\top\boldsymbol{q} & \bar{V}_{\mathcal{I},N_0}= \max & (\bar{\boldsymbol{r}}_{N_0})^\top\boldsymbol{q}\\ \text{s.t.} & C\boldsymbol{q}\le\boldsymbol{\alpha} & \text{s.t.} & \bar{C}_{N_0}\boldsymbol{q}\le\boldsymbol{\alpha}+\lambda_{N_0}\\ & B\boldsymbol{q}=\boldsymbol{\mu} & & |\bar{B}_{N_0}\boldsymbol{q}-\boldsymbol{\mu}|\le\lambda_{N_0}\\ & \boldsymbol{q}_{\mathcal{I}^c}=0 & & \boldsymbol{q}_{\mathcal{I}^c}=0\\ & \boldsymbol{q}\ge 0, & & \boldsymbol{q}\ge 0. \end{array} \tag{15}$$

where $\mathcal{I}^c$ denotes the complementary set of $\mathcal{I}$. Note that if $V-V_{\mathcal{I}}>0$, we know that $\mathcal{I}^c$ contains a non-zero basic variable. The steps 4-8 in Algorithm 1 reflect this point. Starting from $\mathcal{I}$ denoting the whole index set, we sequentially delete one element i (denoting (s,a) in the infinite horizon discounted problem) from the set $\mathcal{I}$. Once we detected that $V-V_{\mathcal{I}\backslash\{i\}}>0$, we know that i is a non-zero basic variable and we add i back into the set $\mathcal{I}$. In this way, we can classify all the basic variables into the set $\mathcal{I}$. However, since we do not know the exact value of V and $V_{\mathcal{I}}$, we use the estimates and compare the value of $\bar{V}_{N_0}$ and $\bar{V}_{\mathcal{I},N_0}$. For this comparison to be valid, the estimation error has to be smaller than the intrinsic gap between V and $V_{\mathcal{I}}$. We define a constant

$$\Delta_1=\min_{\mathcal{I}}\{V-V_{\mathcal{I}}:V-V_{\mathcal{I}}>0\} \tag{16}$$

and we need N_0 to be large enough such that the estimation gap is smaller than $\Delta_1/2$.

To find the corresponding active constraints, we consider the dual program of $V_{\mathcal{I}}$, where $\mathcal{I}$ is determined in steps 4-8 in Algorithm 1, and similarly, we test which dual variable can be set to 0 without influencing the dual objective value. For a dual variable index subset $\mathcal{J} \subset [K]$, we consider the dual program as follows.

$$\begin{aligned} \text{Dual}_{\mathcal{J},\mathcal{I}} = \min \quad & \boldsymbol{\alpha}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{z} && \text{(17a)} \\ \text{s.t.} \quad & (C(:,\mathcal{I}))^\top \boldsymbol{y} + (B(:,\mathcal{I}))^\top \boldsymbol{z} \geq \hat{\boldsymbol{r}}_{\mathcal{I}} && \text{(17b)} \\ & \boldsymbol{y}_{\mathcal{J}^c} = 0 && \text{(17c)} \\ & \boldsymbol{y} \geq 0, \boldsymbol{z} \geq -\infty, && \text{(17d)} \end{aligned}$$

where $\mathcal{J}^c = [K] \backslash \mathcal{J}$. The estimate of $\text{Dual}_{\mathcal{J},\mathcal{I}}$, can be obtained from the estimate of its dual, which is given below.

$$\begin{aligned} \text{Dual}_{\mathcal{J},\mathcal{I},N_0} = \bar{V}_{\mathcal{J},\mathcal{I},N_0} = \max \quad & (\bar{\boldsymbol{r}}_{N_0})^\top \boldsymbol{q} && \text{(18a)} \\ \text{s.t.} \quad & \bar{C}_{N_0}(\mathcal{J},:)\boldsymbol{q} \leq \boldsymbol{\alpha} + \lambda_{N_0} && \text{(18b)} \\ & |\bar{B}_{N_0}\boldsymbol{q} - \boldsymbol{\mu}| \leq \lambda_{N_0} && \text{(18c)} \\ & \boldsymbol{q}_{\mathcal{I}} = 0 && \text{(18d)} \\ & \boldsymbol{q} \geq 0, && \text{(18e)} \end{aligned}$$

Similarly, we compare the value of $\text{Dual}_{\mathcal{I}}$, where $\text{Dual}_{\mathcal{I}} = \text{Dual}_{\mathcal{J}',\mathcal{I}}$ with $\mathcal{J}' = [K]$, and $\text{Dual}_{\mathcal{J},\mathcal{I}}$. However, we can only compare the value of their estimates $\text{Dual}_{\mathcal{I},N_0}$ and $\text{Dual}_{\mathcal{J},\mathcal{I},N_0}$. To this end, we define a constant

$$\Delta_2 = \min_{\mathcal{J} \subset [K]} \{\text{Dual}_{\mathcal{J},\mathcal{I}} - \text{Dual}_{\mathcal{I}} : \text{Dual}_{\mathcal{J},\mathcal{I}} - \text{Dual}_{\mathcal{I}} > 0\}. \tag{19}$$

For the comparison to be valid, we need N_0 to be large enough such that the estimation error is smaller than $\Delta_2/2$. In this way, we identify the linearly independent binding constraints corresponding to $\boldsymbol{q}^*$, as described in steps 9-13 of Algorithm 1.

4 Our Final Algorithm

We now describe our formal algorithm. From the output of Algorithm 1, we characterize one optimal solution. If the sample size n is used in Algorithm 1, we denote by $\mathcal{I}_n$ and $\mathcal{J}_n$ the output of Algorithm 1. To be specific, we have $\boldsymbol{q}^*_{\mathcal{I}_n^c} = 0$ and the non-zero elements $\boldsymbol{q}^*_{\mathcal{I}_n}$ can be given as the solution to

$$\begin{bmatrix} C(\mathcal{J}_n, \mathcal{I}_n) \\ B(:, \mathcal{I}_n) \end{bmatrix} \cdot \boldsymbol{q}^*_{\mathcal{I}_n} = \begin{bmatrix} \boldsymbol{\alpha}_{\mathcal{J}_n} \\ (1-\gamma) \cdot \boldsymbol{\mu} \end{bmatrix}. \tag{20}$$

However, in practice, both the matrices $C(\mathcal{J}_n, \mathcal{I}_n)$ and $B(:, \mathcal{I}_n)$ are unknown. We aim to use the samples to learn the matrices $C(\mathcal{J}_n, \mathcal{I}_n)$ and $B(:, \mathcal{I}_n)$ such that the $\boldsymbol{q}^*_{\mathcal{I}_n}$ can also be determined. Our formal algorithm is given in Algorithm 2. The steps 3-8 in Algorithm 2 is to use Algorithm 1 as a subroutine to identify the set $\mathcal{I}$ and $\mathcal{J}$ that satisfy the conditions in Theorem 2.1. We can show that as long as $n \geq N_0'$, where N_0' is a threshold that depends on the problem parameters, Algorithm 1 correctly obtains the set $\mathcal{I}$ and $\mathcal{J}$ satisfying the conditions in Theorem 2.1. Therefore, we exponentially increase the value of n as input to Algorithm 1 to reach N_0'.

A crucial element in Algorithm 2 (step 10) is that we adaptively update the value of $\boldsymbol{\alpha}^n_{\mathcal{J}_{n-1}}$ and $\boldsymbol{\mu}^n(s')$ as in (22) and (23). We then use the updated $\boldsymbol{\alpha}^n_{\mathcal{J}_{n-1}}$ and $\boldsymbol{\mu}^n(s')$ to obtain the value of $\tilde{\boldsymbol{q}}^n_{\mathcal{I}_n}$ as in (21). Such an algorithmic design follows the resolving idea from online LP to achieve a problem-dependent bound. In step 11, we further project $\tilde{\boldsymbol{q}}^{n+1}$ to the set $\{\boldsymbol{q} \geq 0 : \|\boldsymbol{q}\|_1 \leq 2\}$ to obtain $\boldsymbol{q}^{n+1}$. This prevents $\tilde{\boldsymbol{q}}^n$ from behaving ill when n is not large and the estimates of $C(\mathcal{J}_n, \mathcal{I}_n)$ and $B(:, \mathcal{I}_n)$ are not accurate enough. We can show that when n is large enough, $\tilde{\boldsymbol{q}}^{n+1}$ automatically belongs to the set $\{\boldsymbol{q} \geq 0 : \|\boldsymbol{q}\|_1 \leq 2\}$.

5 Theoretical Analysis

In this section, we conduct our theoretical analysis. The analysis can be divided into two parts. In the first part, we show that Algorithm 1 can successfully help us identify one optimal basis to work

Algorithm 2 The Adaptive-resolving Algorithm

Input: the number of samples N for each $(s,a) \in \mathcal{S} \times \mathcal{A}$.
Initialize $\mathcal{F}_1 = \emptyset$, $\boldsymbol{\alpha}^1 = N \cdot \boldsymbol{\alpha}$ and $\boldsymbol{\mu}^1 = N \cdot \boldsymbol{\mu}$.
for $n = 1, \ldots, N$ **do**
  if $n = 2^m$ for an integer m **then**
    Obtain the output $\mathcal{I}_n$ and $\mathcal{J}_n$ from Algorithm 1 with the input $\mathcal{F}_n$.
  else
    set $\mathcal{I}_n = \mathcal{I}_{n-1}$ and $\mathcal{J}_n = \mathcal{J}_{n-1}$.
  end if
  Construct estimates $\bar{C}^n(\mathcal{J}_n, \mathcal{I}_n)$ and $\bar{B}^n(:, \mathcal{I}_n)$ using the sample set $\mathcal{F}_n$.
  Construct a solution $\tilde{\boldsymbol{q}}^n$ such that $\tilde{\boldsymbol{q}}^n_{\mathcal{I}_n^c} = 0$ and $\tilde{\boldsymbol{q}}^*_{\mathcal{I}_n}$ is the solution to

$$\begin{bmatrix} \bar{C}^n(\mathcal{J}_n, \mathcal{I}_n) \\ \bar{B}^n(:, \mathcal{I}_n) \end{bmatrix} \cdot \tilde{\boldsymbol{q}}^n_{\mathcal{I}_n} = \begin{bmatrix} \dfrac{\boldsymbol{\alpha}^n_{\mathcal{J}_n}}{N-n+1} \\ \dfrac{\boldsymbol{\mu}^n}{N-n+1} \end{bmatrix}. \tag{21}$$

  Project $\tilde{\boldsymbol{q}}^n$ to the set $\{\boldsymbol{q} : \|\boldsymbol{q}\|_1 \le 2\}$ to obtain $\boldsymbol{q}^n$.
  For each $(s,a) \in \mathcal{I}_n \subset \mathcal{S} \times \mathcal{A}$, we query the model $\mathcal{M}$ to obtain a sample of the reward $r^n(s,a)$ and the costs $c^n_k(s,a)$ for each $k \in \mathcal{J}_n \subset [K]$, as well as the state transition $s^n(s,a)$.
  Update $\mathcal{F}_{n+1} = \mathcal{F}_n \cup \{r^n(s,a), c^n_k(s,a), s^n(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \forall k \in [K]\}$.
  Denote by $\boldsymbol{c}^n(s,a) = (c^n_k(s,a))_{\forall k \in \mathcal{J}_n}$ and do the update:

$$\boldsymbol{\alpha}^{n+1}_{\mathcal{J}_n} = \boldsymbol{\alpha}^n_{\mathcal{J}_n} - \sum_{(s,a) \in \mathcal{I}_n} \boldsymbol{c}^n(s,a) \cdot q^n(s,a). \tag{22}$$

  Do the update:

$$\boldsymbol{\mu}^{n+1}(s') = \boldsymbol{\mu}^n(s') - \sum_{(s,a) \in \mathcal{I}} q^n(s,a) \cdot (\delta_{s',s} - \gamma \mathbb{1}_{\{s' = s^n(s,a)\}}), \quad \forall s' \in \mathcal{S}. \tag{23}$$

  where $\mathbb{1}_{\{s' = s^n(s,a)\}}$ is an indicator function of whether the state transition $s^n(s,a)$ equals s'.
end for
We define $\bar{\boldsymbol{q}}^N$ such that $\bar{\boldsymbol{q}}^N_{\mathcal{I}_N^c} = 0$ and $\bar{\boldsymbol{q}}^N_{\mathcal{I}_N} = \frac{1}{N} \cdot \sum_{n=1}^N \boldsymbol{q}^n_{\mathcal{I}_N}$. We then define a policy $\bar{\pi}^N$

$$P(\bar{\pi}^N(s) = a) = \begin{cases} \dfrac{\bar{q}^N(s,a)}{\sum_{a' \in \mathcal{A}} \bar{q}^N(s,a')}, & \text{if } \sum\limits_{a' \in \mathcal{A}} \bar{q}^N(s,a') > 0 \\ 1/|\mathcal{A}|, & \text{if } \sum\limits_{a' \in \mathcal{A}} \bar{q}^N(s,a') = 0. \end{cases} \tag{24}$$

Output: policy $\bar{\pi}^N$.

with. In the second part, we show how to learn the optimal distribution over the optimal basis we have identified.

We now present the theorem showing that Algorithm 1 indeed helps us identify one optimal basis with a high probability. In practice, the value of ε will be set to $1/N$ in the following theorem.

Theorem 5.1 *For any $\varepsilon > 0$, as long as N_0 satisfies the condition*

$$2\mathrm{Gap}_1(N_0, \varepsilon) + 2\mathrm{Gap}_2(N_0, \varepsilon) \le \min\{\Delta_1, \Delta_2\} \tag{25}$$

the outputs $\mathcal{I}_{N_0}$ and $\mathcal{J}_{N_0}$ of Algorithm 1 satisfy the conditions described in Lemma 2.1 with probability at least $1 - (K|\mathcal{S}||\mathcal{A}| - |\mathcal{S}|^2|\mathcal{A}|) \cdot \varepsilon$. Moreover, the sets $\mathcal{I}_n$ and $\mathcal{J}_n$ will be common for any $n \ge N_0$ satisfying (25), which we denote by $\mathcal{I}^*$ and $\mathcal{J}^*$.

An important problem parameter related to $\mathcal{I}^*$ and $\mathcal{J}^*$ can be described as follows. Define

$$A^* = \begin{bmatrix} C(\mathcal{J}^*, \mathcal{I}^*) \\ B(:, \mathcal{I}^*) \end{bmatrix}. \tag{26}$$

We then denote by $\{\sigma_1(A^*), \ldots, \sigma_{|\mathcal{S}|+K'}(A^*)\}$ the eigenvalues of the matrix A^*. We define σ as

$$\sigma = \min\left\{|\sigma_1(A^*)|, \ldots, |\sigma_{|\mathcal{S}|+K'}(A^*)|\right\}. \tag{27}$$

From the non-singularity of the matrix A^*, we know that $\sigma > 0$. We then have the following bound.

Theorem 5.2 *With a sample complexity bound of*

$$O\left(\frac{(|\mathcal{S}|+K)^3 \cdot |\mathcal{S}| \cdot |\mathcal{A}|}{\alpha^2 \cdot \xi \cdot \sigma(1-\gamma) \cdot \min\{\sigma^2, (1-\gamma)^2 \cdot \Delta\}} \cdot \frac{\log^2(1/\varepsilon)}{\varepsilon}\right),$$

where $\Delta = \min\{\Delta_1^2, \Delta_2^2\}$, $\xi = \min_{(s,a)\in\mathcal{I}^*}\{q^*(s,a)\}$ *and* $\boldsymbol{q}^*$ *denotes the optimal soluton to LP (6) corresponding to the optimal basis* $\mathcal{I}^*$ *and* $\mathcal{J}^*$*, we obtain a policy* $\bar{\pi}^N$ *from Algorithm 2 (defined in (24)) such that*

$$V_r(\pi^*, \mu_1) - V_r(\bar{\pi}^N, \mu_1) \leq \varepsilon \;\; \textit{and} \;\; V_k(\bar{\pi}^N, \mu_1) - \alpha_k \leq \varepsilon, \; \forall k \in [K].$$

Our Algorithm 2 can be directly applied to solving MDP problems without resource constraints and our bounds in Theorem E.2 and Theorem 5.2 still hold. Note that in the MDP problems, the parameter σ can be lower bounded by $1-\gamma$, which follows from the fact that the matrix A^* can simply be represented by the probability transition matrix. Then, we have the following sample complexity bound for our Algorithm 2.

Proposition 5.3 *For the MDP problems without resource constraints, i.e.,* $K = 0$*, with a sample complexity bound of*

$$O\left(\frac{|\mathcal{S}|^4 \cdot |\mathcal{A}|}{(1-\gamma)^4 \cdot \xi \cdot \Delta} \cdot \frac{\log^2(1/\varepsilon)}{\varepsilon}\right), \tag{28}$$

we obtain a policy $\bar{\pi}^N$ *from Algorithm 2 (defined in (24)) such that*

$$V_r(\pi^*, \mu_1) - V_r(\bar{\pi}^N, \mu_1) \leq \varepsilon \;\; \textit{and} \;\; V_k(\bar{\pi}^N, \mu_1) - \alpha_k \leq \varepsilon, \; \forall k \in [K].$$

In terms of the dependency of our sample complexity bound on other problem parameters such as $|\mathcal{S}|$, $|\mathcal{A}|$, and $1-\gamma$, we compare to the series of work [58, 67, 68, 2, 27], that subsequently achieves a sample complexity bound of $O\left(\frac{|\mathcal{S}|\cdot|\mathcal{A}|}{(1-\gamma)^3\cdot\varepsilon^2}\right)$, where the dependency over $|\mathcal{S}|$, $|\mathcal{A}|$, and $1-\gamma$ is optimal [24, 43]. Our sample complexity bound in (28) has a worse dependency in terms of $|\mathcal{S}|$ and $1-\gamma$. This is because we construct an empirical LP to estimate the value of LP (6) and resolve the linear equation as in (21), where the size of the LP (which is $|\mathcal{S}|$) and the eigenvalues of the matrix A^* (which is bounded by $1-\gamma$) will play a part. However, our bound (28) enjoys a better dependency in terms of ε.

6 Conclusions

In this paper, we develop the first instance-dependent $\tilde{O}(1/\epsilon)$ sample complexity for constrained MDP problems. We characterize the instance hardness via corner points of the LP formulation and we develop a resolving algorithm to learn the optimal solution while sticking to the identified optimal basis. The work presented by this paper advances the field of Machine Learning and the algorithmic ideas developed in this paper have a broader impact to inspire new algorithms. Our results are developed for the tabular settings, which pose some limitations to the real-world applications of our methods. We leave the extensions to more involved settings for future work.

Acknoledgement

Jiashuo Jiang is generously supported by the early career scheme 26210223 and the general research fund 16204024 from Research Grants Council, Hong Kong.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.

[2] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

[3] Shipra Agrawal, Zizhuo Wang, and Yinyu Ye. A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62(4):876–890, 2014.

[4] Eitan Altman. *Constrained Markov decision processes*. vol7. CRCPress, 1999.

[5] Alessandro Arlotto and Itai Gurvich. Uniformly bounded regret in the multisecretary problem. *Stochastic Systems*, 9(3):231–260, 2019.

[6] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[7] Mohammad Gheshlaghi Azar, Rémi Munos, Mohammad Ghavamzadeh, and Hilbert Kappen. Reinforcement learning with a near optimal rate of convergence. 2011.

[8] Qinbo Bai, Amrit Singh Bedi, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via conservative natural policy gradient primal-dual algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6737–6744, 2023.

[9] Santiago Balseiro, Haihao Lu, and Vahab Mirrokni. Dual mirror descent for online allocation problems. In *International Conference on Machine Learning*, pages 613–628. PMLR, 2020.

[10] Pornpawee Bumpensanti and He Wang. A re-solving heuristic with uniformly bounded loss for network revenue management. *Management Science*, 66(7):2993–3009, 2020.

[11] Boxiao Chen, Jiashuo Jiang, Jiawei Zhang, and Zhengyuan Zhou. Learning to order for inventory systems with lost sales and uncertain supplies. *Management Science*, 2024.

[12] Yi Chen, Jing Dong, and Zhaoran Wang. A primal-dual approach to constrained markov decision processes. *arXiv preprint arXiv:2101.10895*, 2021.

[13] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2018.

[14] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

[15] Yinlam Chow, Ofir Nachum, Aleksandra Faust, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*, 2019.

[16] Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

[17] Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:1–12, 2021.

[18] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

[19] Dongsheng Ding, Kaiqing Zhang, Jiali Duan, Tamer Başar, and Mihailo R Jovanović. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained mdps. *arXiv preprint arXiv:2206.02346*, 2022.

[20] Yaqi Duan and Martin J Wainwright. Taming" data-hungry" reinforcement learning? stability in continuous state-action spaces. *arXiv preprint arXiv:2401.05233*, 2024.

[21] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

[22] Jaime F Fisac, Anayo K Akametalu, Melanie N Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

[23] Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

[24] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91:325–349, 2013.

[25] Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Achieving sub-linear regret in infinite horizon average reward constrained mdp with linear function approximation. In *The Eleventh International Conference on Learning Representations*, 2022.

[26] Egor Gladin, Maksim Lavrik-Karmazin, Karina Zainullina, Varvara Rudenko, Alexander Gasnikov, and Martin Takac. Algorithm for constrained markov decision process with linear convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 11506–11533. PMLR, 2023.

[27] Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted mdps. *Advances in Neural Information Processing Systems*, 34:22288–22300, 2021.

[28] Nicholas J Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2002.

[29] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*, volume 305. Springer science & business media, 1996.

[30] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[31] Stefanus Jasin and Sunil Kumar. A re-solving heuristic with bounded revenue loss for network revenue management with customer choice. *Mathematics of Operations Research*, 37(2):313–345, 2012.

[32] Stefanus Jasin and Sunil Kumar. Analysis of deterministic lp-based booking limit and bid price controls for revenue management. *Operations Research*, 61(6):1312–1320, 2013.

[33] Jiashuo Jiang. Constant approximation for network revenue management with markovian-correlated customer arrivals. *arXiv preprint arXiv:2305.05829*, 2023.

[34] Jiashuo Jiang, Xiaocheng Li, and Jiawei Zhang. Online stochastic optimization with wasserstein based non-stationarity. *arXiv preprint arXiv:2012.06961*, 2020.

[35] Jiashuo Jiang, Will Ma, and Jiawei Zhang. Degeneracy is ok: Logarithmic regret for network revenue management with indiscrete distributions. *arXiv preprint arXiv:2210.07996*, 2022.

[36] Jiashuo Jiang and Jiawei Zhang. Online resource allocation with stochastic resource consumption. *arXiv preprint arXiv:2012.07933*, 2020.

[37] Yujia Jin and Aaron Sidford. Efficiently solving mdps with stochastic mirror descent. In *International Conference on Machine Learning*, pages 4890–4900. PMLR, 2020.

[38] Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

[39] Michael Kearns, Yishay Mansour, and Andrew Y Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine learning*, 49:193–208, 2002.

[40] Thomas Kesselheim, Andreas Tönnis, Klaus Radke, and Berthold Vöcking. Primal beats dual on online packing lps in the random-order model. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 303–312, 2014.

[41] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[42] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[43] Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*, pages 320–334. Springer, 2012.

[44] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*, 72(1):203–221, 2024.

[45] Tianjiao Li, Ziwei Guan, Shaofeng Zou, Tengyu Xu, Yingbin Liang, and Guanghui Lan. Faster algorithm and sharper analysis for constrained markov decision process. *arXiv preprint arXiv:2110.10351*, 2021.

[46] Xiaocheng Li, Chunlin Sun, and Yinyu Ye. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In *International Conference on Machine Learning*, pages 6483–6492. PMLR, 2021.

[47] Xiaocheng Li and Yinyu Ye. Online linear programming: Dual convergence, new algorithms, and regret bounds. *Operations Research*, 70(5):2948–2966, 2022.

[48] Shang Liu, Jiashuo Jiang, and Xiaocheng Li. Non-stationary bandits with knapsacks. *Advances in Neural Information Processing Systems*, 35:16522–16532, 2022.

[49] Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021.

[50] Tao Liu, Ruida Zhou, Dileep Kalathil, PR Kumar, and Chao Tian. Policy optimization for constrained mdps with provable fast global convergence. *arXiv preprint arXiv:2111.00552*, 2021.

[51] Yongshuai Liu, Jiaxin Ding, and Xin Liu. Ipo: Interior-point policy optimization under constraints. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4940–4947, 2020.

[52] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[53] Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, and Tom Zahavy. Reload: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained mdps. In *International Conference on Machine Learning*, pages 25303–25336. PMLR, 2023.

[54] Wenlong Mou, Zheng Wen, and Xi Chen. On the sample complexity of reinforcement learning with policy space generalization. *arXiv preprint arXiv:2008.07353*, 2020.

[55] Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142:205–228, 2009.

[56] Masahiro Ono, Marco Pavone, Yoshiaki Kuwata, and J Balaram. Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Autonomous Robots*, 39:555–571, 2015.

[57] Guy Shani, David Heckerman, Ronen I Brafman, and Craig Boutilier. An mdp-based recommender system. *Journal of Machine Learning Research*, 6(9), 2005.

[58] Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, 31, 2018.

[59] Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

[60] Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pages 9133–9143. PMLR, 2020.

[61] Zhongchang Sun, Sihong He, Fei Miao, and Shaofeng Zou. Constrained reinforcement learning under model mismatch. *arXiv preprint arXiv:2405.01327*, 2024.

[62] Aviv Tamar, Dotan Di Castro, and Shie Mannor. Policy gradients with variance related risk criteria. In *Proceedings of the twenty-ninth international conference on machine learning*, pages 387–396, 2012.

[63] Sharan Vaswani, Lin Yang, and Csaba Szepesvári. Near-optimal sample complexity bounds for constrained mdps. *Advances in Neural Information Processing Systems*, 35:3110–3122, 2022.

[64] Alberto Vera and Siddhartha Banerjee. The bayesian prophet: A low-regret framework for online decision making. *Management Science*, 67(3):1368–1391, 2021.

[65] Andrew J Wagenmaker and Dylan J Foster. Instance-optimality in interactive decision making: Toward a non-asymptotic theory. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1322–1472. PMLR, 2023.

[66] Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning. In *Conference on Learning Theory*, pages 358–418. PMLR, 2022.

[67] Martin J Wainwright. Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

[68] Mengdi Wang. Randomized linear programming solves the markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45(2):517–546, 2020.

[69] Honghao Wei, Xin Liu, and Lei Ying. A provably-efficient model-free algorithm for constrained markov decision processes. *arXiv preprint arXiv:2106.01577*, 2021.

[70] Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR, 2021.

[71] Yunbei Xu and Assaf Zeevi. Towards optimal problem dependent generalization error bounds in statistical learning theory. *Mathematics of Operations Research*, 2024.

[72] Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

[73] Sihan Zeng, Thinh T Doan, and Justin Romberg. Finite-time complexity of online primal-dual natural actor-critic algorithm for constrained markov decision processes. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 4028–4033. IEEE, 2022.

[74] Ruida Zhou, Tao Liu, Dileep Kalathil, PR Kumar, and Chao Tian. Anchor-changing regularized natural policy gradient for multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 35:13584–13596, 2022.

A Numerical Experiments

We implement our Algorithm 2 to study the numerical performance. We consider a CMDP problem with the state space $|\mathcal{S}| = 10$ and the action space $|\mathcal{A}| = 10$. We set the discount factor $\gamma = 0.7$. We then randomly generate the probability transition kernel P. To be specific, for each state $s \in \mathcal{S}$, action $a \in \mathcal{A}$, and the future state $s' \in \mathcal{S}$, we uniformly generate a randomly variable $p_{s,a,s'}$. Then, the transition probability is defined as $P(s'|s,a) = \frac{p_{s,a,s'}}{\sum_{s'' \in \mathcal{S}} p_{s,a,s''}}$. For each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, the expected reward $\hat{r}(s,a)$ is uniformly generated from the interval $[1,2]$ (with the reward for the first action set to be 0). The actual reward $r(s,a) = \hat{r}(s,a) + \eta$, where η is uniformly distributed among $[-0.5, 0.5]$. There are $K = 5$ constraints and for each constraint $k \in [K]$ and each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we define the expected cost $\hat{c}_k(s,a)$ to be uniformly generated from $[1,2]$. The actual cost $c_k(s,a) = \hat{c}_k(s,a) + \eta'$, where η' is uniformly distributed among $[-0.5, 0.5]$.

For each total iterations N, We apply Algorithm 2 and obtain the output $\boldsymbol{q}^1, \ldots, \boldsymbol{q}^N$. We compare $\bar{\boldsymbol{q}}^N$ with the optimal occupancy measure. Since our algorithm is a randomized algorithm, we study the performance of our algorithm in expectation. To be specific, given the problem instance and a fixed N, we implement our algorithm repeatedly for $M = 500$ rounds. Denote by $\bar{\boldsymbol{q}}_m^N$ the output of our Algorithm 2 at round m, for $m \in [M]$. We define the error term as $\text{Err}(N) = \frac{1}{M} \cdot \sum_{m=1}^{M} \|\bar{\boldsymbol{q}}_m^N - \boldsymbol{q}^*\|_1 / \|\boldsymbol{q}^*\|_1$. We study how the error term $\text{Err}(N)$ scales with N. The results are displayed in Figure 2. As we can see, the error term $\text{Err}(N)$ decreases quickly with respect to N. Moreover, since our Algorithm 2 requires only solving a set of linear equations in each iteration, the computation cost of our Algorithm 2 is also moderate.

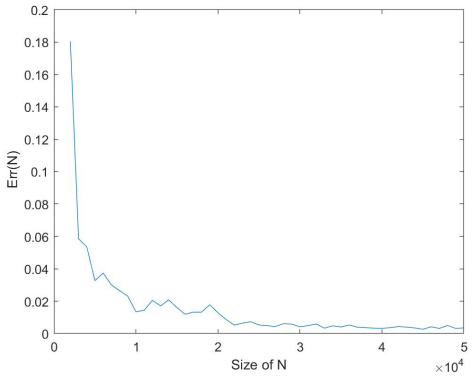

Figure 2: The computational performance of Algorithm 2. The x-label denotes the size of N, while the y-label denotes the error term $\text{Err}(N)$.

B Proof of Lemma 2.1

Let $\boldsymbol{q}^*$ be a basic optimal solution to LP (6). We also denote by $\mathcal{I}$ the index set of $\boldsymbol{q}^*$ such that $\boldsymbol{q}^*_{\mathcal{I}} > 0$. We then denote by $\mathcal{J}' \in [K]$ the set of active constraints at $\boldsymbol{q}^*$. Then, the following linear equations must be satisfied by $\boldsymbol{q}^*$.

$$C(\mathcal{J}', \mathcal{I}) \boldsymbol{q}^*_{\mathcal{I}} = \boldsymbol{\alpha}_{\mathcal{J}'} \text{ and } B(:, \mathcal{I}) \boldsymbol{q}^*_{\mathcal{I}} = \boldsymbol{\mu}. \tag{29}$$

From the property of the basic optimal solution, we must have $|\mathcal{J}'| + b \geq |\mathcal{I}|$.

If $|\mathcal{J}'| + b = |\mathcal{I}|$, then we know that we already find the desired index set $\mathcal{I}$ and $\mathcal{J} = \mathcal{J}'$.

Otherwise, if $|\mathcal{J}'| + b > |\mathcal{I}|$, then we know that the LP (6) is degenerate. However, now the linear system (29) is over-determined, i.e., there must be $|\mathcal{J}'| + b - |\mathcal{I}|$ number of equations can be implied by the others. It only remains to show that those redundant equations are all in the set $\mathcal{J}'$. This step can be done can showing that the matrix $B(:, \mathcal{I})$ has a full row rank. We prove this by showing contradiction.

For the infinite horizon discounted problem, suppose that $B(:,\mathcal{I})$ does not have full row rank, there must exists $\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{S}|}$ such that

$$\boldsymbol{\beta}^\top B(:,\mathcal{I}) = \mathbf{0}. \tag{30}$$

On the other hand, denote by π^* the optimal policy corresponding to $\boldsymbol{q}^*$. We then define a matrix $\Gamma \in \mathbb{R}^{|\mathcal{I}|\times\mathcal{S}}$ such that the s-th column of Γ is a vector that takes 0 for all the $(s', a) \in \mathcal{I}$-th element if $s' \neq s$, and take a value of $\pi^*(a|s)$ for all a such that $(s, a) \in \mathcal{I}$. We know that

$$B(:,\mathcal{I})\Gamma = I - \gamma \cdot P^{\pi^*}$$

with $I \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ being the identify matrix and P^{π^*} being the transition probability matrix under the policy π^*, with the element at the s-th row and s'-th column denoting the probability of transiting from state s to state s' under the policy π^*. It is well known that the matrix $I - \gamma \cdot P^{\pi^*}$ is non-singular (see for example the Gersgorin's Theorem in [30]). Therefore, for any vector $\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{S}|}$, we know that

$$\boldsymbol{\beta}^\top B(:,\mathcal{I})\Gamma = \boldsymbol{\beta}^\top (I - \gamma \cdot P^{\pi^*}) \neq 0$$

which contradicts with (60). Our proof is thus completed.

C Proof of Lemma 3.2

We now condition on the event that $|\bar{r}_{N_0}(s,a) - \hat{r}(s,a)|$, $|\bar{c}_{k,N_0}(s,a) - \hat{c}_k(s,a)|$, and $|\bar{P}_{N_0}(s'|s,a) - P(s'|s,a)|$ for each $k \in [K]$, $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $s' \in \mathcal{S}$ are all bounded by $\mathrm{Rad}(N_0, \varepsilon)$. From the union bound, we know that this event happens with probability at least $1 - (K|\mathcal{S}||\mathcal{A}| - |\mathcal{S}|^2|\mathcal{A}|)\cdot\varepsilon$.

Note that for LP (4), by summing up the constraint (4c) for all $s \in \mathcal{S}$, we obtain that any feasible solution $\boldsymbol{q}$ for LP (4) would satisfy

$$\|\boldsymbol{q}\|_1 = 1. \tag{31}$$

We first upper bound the gap $V - \bar{V}_{N_0}$. Denote by $\boldsymbol{q}^*$ one optimal solution to V. Then, from the feasibility of $\boldsymbol{q}^*$, we know that

$$\bar{C}_{N_0}\boldsymbol{q}^* = C\boldsymbol{q}^* + (\bar{C}_{N_0} - C)\boldsymbol{q}^* \leq \boldsymbol{\alpha} + \mathrm{Rad}(N_0,\varepsilon) \leq \boldsymbol{\alpha} + \lambda_{N_0} \tag{32}$$

where the first inequality follows from $\|\boldsymbol{q}^*\|_1 = 1$ (31) and all elements of $\bar{C}_{N_0} - C$ are upper bounded by $\mathrm{Rad}(N_0,\varepsilon)$. Also, we know that

$$\bar{B}_{N_0}\boldsymbol{q}^* = B\boldsymbol{q}^* + (\bar{B}_{N_0} - B)\boldsymbol{q}^* \leq (1-\gamma)\cdot\boldsymbol{\mu} + \gamma\cdot\mathrm{Rad}(N_0,\varepsilon) \leq (1-\gamma)\cdot\boldsymbol{\mu} + \lambda_{N_0} \tag{33}$$

where the first inequality follows from $\|\boldsymbol{q}^*\|_1 = 1$ (31) and all elements of $\bar{B}_{N_0} - B$ are upper bounded by $\gamma\cdot\mathrm{Rad}(N_0,\varepsilon)$. Similarly, we have that

$$\bar{B}_{N_0}\boldsymbol{q}^* = B\boldsymbol{q}^* + (\bar{B}_{N_0} - B)\boldsymbol{q}^* \geq (1-\gamma)\cdot\boldsymbol{\mu} - \gamma\cdot\mathrm{Rad}(N_0,\varepsilon) \geq (1-\gamma)\cdot\boldsymbol{\mu} - \lambda_{N_0}. \tag{34}$$

Therefore, as long as

$$\lambda_{N_0} \geq \mathrm{Rad}(N_0,\varepsilon), \tag{35}$$

we know that $\boldsymbol{q}^*$ is a feasible solution to $\bar{V}_{N_0}$. We have that

$$\bar{\boldsymbol{r}}_{N_0}^\top\boldsymbol{q}^* \geq \hat{\boldsymbol{r}}^\top\boldsymbol{q}^* - \mathrm{Rad}(N_0,\varepsilon) \tag{36}$$

by noting $\|\boldsymbol{q}^*\|_1 = 1$ (31) and all elements of $\bar{\boldsymbol{r}}_{N_0} - \hat{\boldsymbol{r}}$ are upper bounded by $\mathrm{Rad}(N_0,\varepsilon)$. Therefore, we can obtain the bound

$$V - \mathrm{Rad}(N_0,\varepsilon) \leq \bar{\boldsymbol{r}}_{N_0}^\top\boldsymbol{q}^* \leq \bar{V}_{N_0}. \tag{37}$$

We then lower bound the gap $V - \bar{V}_{N_0}$. We first define

$$\begin{aligned}
\bar{V}_{N_0}(\lambda'_{N_0}) = \max \quad & (\bar{\boldsymbol{r}}_{N_0} - \lambda'_{N_0})^\top\boldsymbol{q} \\
\text{s.t.} \quad & \bar{C}_{N_0}\boldsymbol{q} \leq \boldsymbol{\alpha} + \lambda_{N_0} \\
& \bar{B}_{N_0}\boldsymbol{q} \leq \boldsymbol{\mu} + \lambda_{N_0} \\
& \bar{B}_{N_0}\boldsymbol{q} \geq \boldsymbol{\mu} - \lambda_{N_0} \\
& \boldsymbol{q} \geq 0,
\end{aligned} \tag{38}$$

for any constant λ'_{N_0}. Clearly, any optimal solution $\bar{\boldsymbol{q}}^*$ to $\bar{V}_{N_0}$ will be a feasible solution to $\bar{V}_{N_0}(\lambda'_{N_0})$. Moreover, by summing up the constraints $\bar{B}_{N_0}\boldsymbol{q} \leq \boldsymbol{\mu} + \lambda_{N_0}$ for all $s \in \mathcal{S}$, we know that

$$\|\bar{\boldsymbol{q}}^*\|_1 \leq 1 + \frac{|\mathcal{S}|}{1-\gamma} \cdot \lambda_{N_0}. \tag{39}$$

Then, it holds that

$$\bar{V}_{N_0} \leq (\bar{\boldsymbol{r}}_{N_0} - \lambda'_{N_0})^\top \bar{\boldsymbol{q}}^* + \lambda'_{N_0} \cdot \|\bar{\boldsymbol{q}}^*\|_1 \leq \bar{V}_{N_0}(\lambda'_{N_0}) + \lambda'_{N_0} + \frac{|\mathcal{S}|}{1-\gamma} \cdot \lambda_{N_0}\lambda'_{N_0}. \tag{40}$$

We then compare the value between $\bar{V}_{N_0}(\lambda'_{N_0})$ and V. The dual of $\bar{V}_{N_0}(\lambda'_{N_0})$ is given below.

$$\begin{aligned} \text{Dual}'_{N_0}(\lambda'_{N_0}) = \min \quad & \boldsymbol{\alpha}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top(\boldsymbol{z}_1 - \boldsymbol{z}_2) + \lambda_{N_0} \cdot (\|\boldsymbol{y}\|_1 + \|\boldsymbol{z}_1\|_1 + \|\boldsymbol{z}_2\|_1) \\ \text{s.t.} \quad & \bar{C}_{N_0}^\top \boldsymbol{y} + \bar{B}_{N_0}^\top(\boldsymbol{z}_1 - \boldsymbol{z}_2) \geq \bar{\boldsymbol{r}}_{N_0} - \lambda'_{N_0} \\ & \boldsymbol{y} \geq 0, \boldsymbol{z}_1 \geq 0, \boldsymbol{z}_2 \geq 0. \end{aligned} \tag{41}$$

Denote by $\boldsymbol{y}^*$ and $\boldsymbol{z}^*$ one optimal solution to the dual of LP (6), given below.

$$\begin{aligned} \text{Dual} = \min \quad & \boldsymbol{\alpha}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{z} \\ \text{s.t.} \quad & C^\top \boldsymbol{y} + B^\top \boldsymbol{z} \geq \hat{\boldsymbol{r}} \\ & \boldsymbol{y} \geq 0, \boldsymbol{z} \geq -\infty. \end{aligned} \tag{42}$$

We now show that $\boldsymbol{y}^*$ and $\boldsymbol{z}^*$ is also a feasible solution to Dual'_{N_0}, with $\lambda'_{N_0} = \frac{1}{\min_{k\in[K]}\{\alpha_k\}} \cdot \left(1 + \frac{|\mathcal{S}|}{1-\gamma}\right) \cdot \text{Rad}(N_0, \varepsilon)$. We have the following claim regarding the upper bound on $\|\boldsymbol{y}^*\|_\infty$ and $\|\boldsymbol{z}^*\|_\infty$.

Claim C.1 *There exists an optimal solution $\boldsymbol{y}^*$ and $\boldsymbol{z}^*$ to the Dual (42) such that*

$$\|\boldsymbol{y}^*\|_1 \leq \frac{1}{\min_{k\in[K]}\{\alpha_k\}} \quad \textit{and} \quad \|\boldsymbol{z}^*\|_\infty \leq \frac{1}{1-\gamma} \cdot \frac{1}{\min_{k\in[K]}\{\alpha_k\}}.$$

Then, we define $\bar{\boldsymbol{y}}^* = \boldsymbol{y}^*$, $\boldsymbol{z}_1^* = \max\{0, \boldsymbol{z}^*\}$ and $\boldsymbol{z}_2^* = \max\{0, -\boldsymbol{z}^*\}$. We have

$$\begin{aligned} \bar{C}_{N_0}^\top \bar{\boldsymbol{y}}^* + \bar{B}_{N_0}^\top(\boldsymbol{z}_1^* - \boldsymbol{z}_2^*) &\geq C^\top \boldsymbol{y}^* + B^\top \boldsymbol{z}^* - \text{Rad}(N_0, \varepsilon) \cdot (\|\boldsymbol{y}^*\|_1 + \|\boldsymbol{z}^*\|_1) \\ &\geq C^\top \boldsymbol{y}^* + B^\top \boldsymbol{z}^* - \text{Rad}(N_0, \varepsilon) \cdot \frac{1}{\min_{k\in[K]}\{\alpha_k\}} \cdot \left(1 + \frac{|\mathcal{S}|}{1-\gamma}\right) \\ &\geq \hat{\boldsymbol{r}} - \frac{1}{\min_{k\in[K]}\{\alpha_k\}} \cdot \left(1 + \frac{|\mathcal{S}|}{1-\gamma}\right) \cdot \text{Rad}(N_0, \varepsilon). \end{aligned}$$

Thus, we know that $\bar{\boldsymbol{y}}^*$ and $\boldsymbol{z}_1^*, \boldsymbol{z}_2^*$ is also a feasible solution to $\text{Dual}_{N_0}(\lambda'_{N_0})$, and we have

$$\text{Dual}'_{N_0}(\lambda'_{N_0}) \leq \text{Dual} + \lambda_{N_0} \cdot (\|\boldsymbol{y}^*\|_1 + \|\boldsymbol{z}^*\|_1) \leq \text{Dual} + \frac{\text{Rad}(N_0, \varepsilon)}{\min_{k\in[K]}\{\alpha_k\}} \cdot \left(1 + \frac{|\mathcal{S}|}{1-\gamma}\right). \tag{43}$$

Combing (43) with (40) and also noting that $\bar{V}_{N_0}(\lambda'_{N_0}) = \text{Dual}'_{N_0}(\lambda'_{N_0})$, we have

$$\bar{V}_{N_0} \leq V + \frac{2\text{Rad}(N_0, \varepsilon)}{\min_{k\in[K]}\{\alpha_k\}} \cdot \left(1 + \frac{|\mathcal{S}|}{1-\gamma}\right) + \frac{\text{Rad}^2(N_0, \varepsilon)}{\min_{k\in[K]}\{\alpha_k\}} \cdot \left(|\mathcal{S}| + \frac{|\mathcal{S}|^2}{1-\gamma}\right)$$

which completes our proof.

Proof of Claim C.1. We first bound $\|\boldsymbol{y}^*\|_\infty$. We utilize the approach in [29, 55]. We define a Lagrangian function, with only $\boldsymbol{y}$ as the Lagrangian dual variable.

$$L(\boldsymbol{y}, \boldsymbol{q}) := \boldsymbol{\alpha}^\top \boldsymbol{y} + \hat{\boldsymbol{r}}^\top \boldsymbol{q} - \boldsymbol{y}^\top C \boldsymbol{q} \tag{44}$$

where the feasible set for $\boldsymbol{q}$ is $\{\boldsymbol{q} \geq 0 : B\boldsymbol{q} = \boldsymbol{\mu}\}$ and the feasible set for $\boldsymbol{y}$ is $\{\boldsymbol{y} \geq 0\}$. Following Lemma 3 in [55], it is without loss of generality to restrict the feasible set of $\boldsymbol{y}$ to the set $\{\boldsymbol{y} \geq 0 : \|\boldsymbol{y}\|_1 \leq \rho\}$, where the constant ρ is defined as

$$\rho = \frac{\hat{\boldsymbol{r}}^\top \boldsymbol{q}^* - \hat{\boldsymbol{r}}^\top \bar{\boldsymbol{q}}}{\min_{k\in[K]}\{\alpha_k - C(k,:)\bar{\boldsymbol{q}}\}} \tag{45}$$

where $\bar{\boldsymbol{q}}$ is the occupancy measure that satisfies Slater's condition as stated in Theorem 3.1. Note that we have

$$\rho \leq \frac{1}{\min_{k\in[K]}\{\alpha_k\}}.$$

Therefore, we obtain the following bound on $\boldsymbol{y}^*$:

$$\|\boldsymbol{y}^*\|_1 \leq \rho. \tag{46}$$

We now proceed to bound $\boldsymbol{z}^*$. Denote by $\boldsymbol{q}^*$ the optimal solution corresponding to the optimal dual solution $(\boldsymbol{y}^*, \boldsymbol{z}^*)$, with $\boldsymbol{y}^*$ bounded as in (46). We also denote by π^* the optimal policy corresponding to $\boldsymbol{q}^*$. Then, from the complementary slackness condition, as long as $\boldsymbol{q}^*(s,a) > 0$ for a (s,a), we must have

$$C(:,(s,a))^\top \boldsymbol{y}^* + B(:,(s,a))^\top \boldsymbol{z}^* = \hat{r}(s,a). \tag{47}$$

We now multiply both sides of (47) by $q^*(s,a)/\sum_{a'\in\mathcal{A}} q^*(s,a')$, and sum up over a, for each state s. Then we get

$$(C^{\pi^*})^\top \boldsymbol{y}^* + B^{\pi^*}\boldsymbol{z}^* = \hat{\boldsymbol{r}}^{\pi^*}. \tag{48}$$

Here, $C^{\pi^*} \in \mathbb{R}^{K\times|\mathcal{S}|}$ and the element at k-th row and s-th column is $\sum_{a\in\mathcal{A}} c_k(s,a)\cdot\frac{q^*(s,a)}{\sum_{a'\in\mathcal{A}} q^*(s,a')}$. $B^{\pi^*} = I - \gamma\cdot P^{\pi^*} \in \mathbb{R}^{|\mathcal{S}|\times\mathcal{S}}$, where $P^{\pi^*}(s,s') = \sum_{a\in\mathcal{A}}\pi^*(a|s)\cdot P(s'|s,a)$ denotes the transition probability matrix under the policy π^*. Also, $\hat{\boldsymbol{r}}^{\pi^*} \in \mathbb{R}^{|\mathcal{S}|}$ with $\hat{r}^{\pi^*}(s) = \sum_{a\in\mathcal{A}}\hat{r}(s,a)\cdot\frac{q^*(s,a)}{\sum_{a'\in\mathcal{A}} q^*(s,a')}$. Then, we have

$$\boldsymbol{z}^* = (B^{\pi^*})^{-1}\cdot\left(\hat{\boldsymbol{r}}^{\pi^*} - (C^{\pi^*})^\top\boldsymbol{y}^*\right) \tag{49}$$

From [37], we know that

$$\|(B^{\pi^*})^{-1}\|_\infty \leq \frac{1}{1-\gamma}. \tag{50}$$

Also, from the bound on $\boldsymbol{y}^*$ in (46), we know that

$$\left\|\hat{\boldsymbol{r}}^{\pi^*} - (C^{\pi^*})^\top\boldsymbol{y}^*\right\|_\infty \leq \rho. \tag{51}$$

Therefore, we have that

$$\|\boldsymbol{z}^*\|_\infty \leq \frac{\rho}{1-\gamma}, \tag{52}$$

which completes our proof. □

D Proof of Theorem 5.1

We now condition on the event that $|\bar{r}_{N_0}(s,a) - \hat{r}(s,a)|$, $|\bar{c}_{k,N_0}(s,a) - \hat{c}_k(s,a)|$, and $|\bar{P}_{N_0}(s'|s,a) - P(s'|s,a)|$ for each $k\in[K]$, $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $s'\in\mathcal{S}$ are all bounded by $\mathrm{Rad}(N_0,\varepsilon)$. We know that this event happens with probability at least $1-(K|\mathcal{S}||\mathcal{A}| - |\mathcal{S}|^2|\mathcal{A}|)\cdot\varepsilon$.

Following the procedure in Algorithm 1, we identify an index set $\mathcal{I}$, where we set $\boldsymbol{q}_i = 0$ for each $i\in\mathcal{I}^c$. Note that we cannot further delete one more i from the set $\mathcal{I}$. Otherwise, when deleting this particular i in $\mathcal{I}$ by setting $\mathcal{I}' = \mathcal{I}\backslash\{i\}$, we will have $V_{\mathcal{I}'} = V$ and as a result of Lemma 3.2, we have

$$|\bar{V}_{N_0} - \bar{V}_{\mathcal{I}',N_0}| \leq |V - V_{\mathcal{I}'}| + 2\mathrm{Gap}_1(N_0,\varepsilon) + 2\mathrm{Gap}_2(N_0,\varepsilon) = 2\mathrm{Gap}_1(N_0,\varepsilon) + 2\mathrm{Gap}_2(N_0,\varepsilon) \tag{53}$$

Therefore, if it is feasible (objective value does not change) to delete one more i from the index set $\mathcal{I}$, our algorithm will already do so.

We denote by $\boldsymbol{q}^*$ the optimal solution to LP (6), corresponding to the optimal basis $\mathcal{I}$ identified by Algorithm 1. We know that $\boldsymbol{q}^*_{\mathcal{I}} > 0$ and $\boldsymbol{q}^*_{\mathcal{I}^c} = 0$, which implies that

$$\begin{aligned} V = V_{\mathcal{I}} = \max \quad & (\hat{\boldsymbol{r}}_{\mathcal{I}})^\top\boldsymbol{q} \\ \text{s.t.} \quad & C(:,\mathcal{I})\boldsymbol{q} \leq \boldsymbol{\alpha} \\ & B(:,\mathcal{I})\boldsymbol{q} = \boldsymbol{\mu} \\ & \boldsymbol{q} \geq 0, \end{aligned} \tag{54}$$

Note that in the formulation of (54), we simply discard the columns of the constraint matrix in the index set $\mathcal{I}$. Thus, one optimal solution to (54) will just be $\boldsymbol{q}^*_{\mathcal{I}}$. The dual of (54) is

$$\begin{aligned} \text{Dual}_{\mathcal{I}} = \min \quad & \boldsymbol{\alpha}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{z} \\ \text{s.t.} \quad & (C(:,\mathcal{I}))^\top \boldsymbol{y} + (B(:,\mathcal{I}))^\top \boldsymbol{z} \geq \hat{\boldsymbol{r}}(\mathcal{I}) \\ & \boldsymbol{y} \geq 0, \boldsymbol{z} \geq -\infty, \end{aligned} \tag{55}$$

From the complementary slackness condition, we know that for the optimal dual variable corresponding to $\boldsymbol{q}^*_{\mathcal{I}}$, all the constraints in (55) must hold with equality. Therefore, we have the following result.

Claim D.1 *It holds that*

$$\begin{aligned} V = V_{\mathcal{I}} = \mathit{Dual}_{\mathcal{I}} = \min \quad & \boldsymbol{\alpha}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{z} \\ \textit{s.t.} \quad & (C(:,\mathcal{I}))^\top \boldsymbol{y} + (B(:,\mathcal{I}))^\top \boldsymbol{z} = \hat{\boldsymbol{r}}(\mathcal{I}) \\ & \boldsymbol{y} \geq 0, \boldsymbol{z} \geq -\infty. \end{aligned} \tag{56}$$

The proof of the claim is relegated to the end of this proof. We now show that Algorithm 1 identify the linearly independent binding constraints for $\boldsymbol{q}^*$ such that the conditions in Lemma 2.1 are satisfied.

Denote by $C(\mathcal{J},\mathcal{I})$ a sub-matrix of C such that the rows in the index set $\mathcal{J} \subset [K]$ and the columns in the index set $\mathcal{I}$ of the matrix C remain. Now if the matrix

$$A = [C(\mathcal{J},\mathcal{I}); B(:,\mathcal{I})] \tag{57}$$

is singular, there must be

(i). one row in $C(\mathcal{J},\mathcal{I})$ can be expressed as the linear combination of the other rows in A (we know that the matrix $B(:,\mathcal{I})$ has full row rank from Lemma 2.1);

(ii). or one column in A can be expressed as the linear combination of other columns.

We consider the situation (i). Denote this row as k' and we know that restricting $y_{k'} = 0$ will not change the objective value of $\text{Dual}_{\mathcal{I}}$, which implies that k' will be added to the index set $\mathcal{J}$. This is because $y_{k'}$ cannot be a basic variable, otherwise, the optimal basis of the Dual (56) will be linearly dependent. Therefore, we obtain a contradiction and we know that the matrix A must have full row rank. Thus, situation (i) will not happen.

We then consider the situation (ii). Note that we have

$$\begin{aligned} V = \text{Dual}_{\mathcal{J},\mathcal{I}} = \min \quad & \boldsymbol{\alpha}_{\mathcal{J}}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{z} \\ \text{s.t.} \quad & (C(\mathcal{J},\mathcal{I}))^\top \boldsymbol{y} + (B(:,\mathcal{I}))^\top \boldsymbol{z} \geq \hat{\boldsymbol{r}}_{\mathcal{I}} \\ & \boldsymbol{y} \geq 0, \boldsymbol{z} \geq -\infty, \end{aligned} \tag{58}$$

Comparing the formulation (58) to the formulation (55), we only remain the rows of the matrix $C(:,\mathcal{I})$ in the index set $\mathcal{J}$, where it is feasible (objective value does not change) to set $\boldsymbol{y}_{\mathcal{J}^c} = 0$. The dual of (58) is

$$\begin{aligned} V = V_{\mathcal{J},\mathcal{I}} = \max \quad & (\hat{\boldsymbol{r}}_{\mathcal{I}})^\top \boldsymbol{q} \\ \text{s.t.} \quad & C(\mathcal{J},\mathcal{I})\boldsymbol{q} \leq \boldsymbol{\alpha}_{\mathcal{J}} \\ & B(:,\mathcal{I})\boldsymbol{q} = \boldsymbol{\mu} \\ & \boldsymbol{q} \geq 0. \end{aligned} \tag{59}$$

If situation (ii) happens and one column of A can be expressed as the linear combination of the other columns in A, then we denote the index of this column by i' and we know that we can simply restrict $q_{i'} = 0$ without changing the objective value of (59). This is because $q_{i'}$ cannot be a basic variable otherwise the corresponding optimal basis will be linearly dependent. However, the index $i' \in \mathcal{I}$. This means that we cannot further delete i' from the index set $\mathcal{I}$ by restricting $q_{i'} = 0$ without changing the objective value of LP (4). The above argument leads to a contradiction. Therefore, we know that situation (ii) cannot happen.

Given the arguments above, we know that the matrix A is a non-singular matrix and thus the conditions in Lemma 2.1 are satisfied with the variable index set $\mathcal{I}$ and the constraint index set $\mathcal{J}$.

It only remains to show that we can tell whether the objective value of V^{Infi} or $\text{Dual}_{\mathcal{I}}$ has changed by restricting one variable to be 0. Following the same steps as in Lemma 3.2, we can show that for index sets $\mathcal{I}'$ and $\mathcal{J}'$, it holds that

$$\text{Gap}_1(N_0, \varepsilon) \geq V_{\mathcal{J}',\mathcal{I}'} - \bar{V}_{\mathcal{J}',\mathcal{I}',N_0} \tag{60}$$

and the upper gap as

$$\text{Gap}_2(N_0, \varepsilon) \geq \bar{V}_{\mathcal{J}',\mathcal{I}',N_0} - V_{\mathcal{J}',\mathcal{I}'} \tag{61}$$

with the formulation of $\text{Gap}_1(N_0, \varepsilon)$ and $\text{Gap}_2(N_0, \varepsilon)$ given in (12) and (13). Therefore, as long as

$$2\text{Gap}_1(N_0, \varepsilon) + 2\text{Gap}_2(N_0, \varepsilon) \geq \min\{\Delta_1, \Delta_2\}, \tag{62}$$

where δ_1 is defined in (16) and δ_2 is defined in (19), we can tell whether V^{Infi} is different from $V^{\text{Infi}}_{\mathcal{I}'}$ and whether $\text{Dual}_{\mathcal{I}'}$ is different from $\text{Dual}_{\mathcal{J}',\mathcal{I}'}$. Our proof is thus completed.

Proof of Claim D.1. We first show that

$$\begin{aligned} V = V_{\mathcal{I}} = \max \quad & (\hat{\boldsymbol{r}}_{\mathcal{I}})^\top \boldsymbol{q} \\ \text{s.t.} \quad & C(:,\mathcal{I})\boldsymbol{q} \leq \boldsymbol{\alpha} \\ & B(:,\mathcal{I})\boldsymbol{q} = \boldsymbol{\mu} \\ & \boldsymbol{q} \geq 0. \end{aligned} \tag{63}$$

Denote by $\boldsymbol{q}^*$ the optimal solution to V corresponding to the optimal basis $\mathcal{I}$. It is clear to see that $\boldsymbol{q}^*_{\mathcal{I}}$ is a feasible solution to $V_{\mathcal{I}}$ with the same objective value. Then, we have

$$V \leq V_{\mathcal{I}}. \tag{64}$$

On the other hand, we denote by $\hat{\boldsymbol{q}}$ one optimal solution to $V_{\mathcal{I}}$, and we construct

$$\tilde{\boldsymbol{q}}_{\mathcal{I}} = \hat{\boldsymbol{q}} \quad \text{and} \quad \tilde{\boldsymbol{q}}_{\mathcal{I}^c} = 0.$$

It is clear to see that $\tilde{\boldsymbol{q}}$ is a feasible solution to V with the same objective value, which implies that

$$V_{\mathcal{I}} \leq V. \tag{65}$$

Therefore, (63) is proved from combining (64) and (65). The dual of $V_{\mathcal{I}}$ is given by

$$\begin{aligned} \text{Dual}_{\mathcal{I}} = \min \quad & \boldsymbol{\alpha}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{z} \\ \text{s.t.} \quad & (C(:,\mathcal{I}))^\top \boldsymbol{y} + (B(:,\mathcal{I}))^\top \boldsymbol{z} \geq \hat{\boldsymbol{r}}(\mathcal{I}) \\ & \boldsymbol{y} \geq 0, \boldsymbol{z} \geq -\infty. \end{aligned} \tag{66}$$

We denote by $\boldsymbol{y}^*, \boldsymbol{z}^*$ the optimal dual variable to V corresponding to $\boldsymbol{q}^*$. It is easy to see that $\boldsymbol{y}^*, \boldsymbol{z}^*$ is also feasible to $\text{Dual}_{\mathcal{I}}$ and

$$\text{Dual}_{\mathcal{I}} = V_{\mathcal{I}} = V = \boldsymbol{\alpha}^\top \boldsymbol{y}^* + \boldsymbol{\mu}^\top \boldsymbol{z}^*$$

where the first equality follows from the strong duality between $V_{\mathcal{I}}$ and $\text{Dual}_{\mathcal{I}}$, the second equality follows from (63), and the third equality follows from the strong duality between V and its dual. Therefore, we know that $\boldsymbol{y}^*$ and $\boldsymbol{z}^*$ is also an optimal solution to $\text{Dual}_{\mathcal{I}}$. Moreover, note that from the complementary slackness condition, since $\boldsymbol{q}^*_{\mathcal{I}} > 0$, we must have

$$(C(:,\mathcal{I}))^\top \boldsymbol{y}^* + (B(:,\mathcal{I}))^\top \boldsymbol{z}^* = \hat{\boldsymbol{r}}(\mathcal{I}). \tag{67}$$

We know that $\boldsymbol{y}^*, \boldsymbol{z}^*$ is a feasible solution to

$$\begin{aligned} \text{Dual}'_{\mathcal{I}} = \min \quad & \boldsymbol{\alpha}^\top \boldsymbol{y} + (1-\gamma) \cdot \boldsymbol{\mu}^\top \boldsymbol{z} \\ \text{s.t.} \quad & (C(:,\mathcal{I}))^\top \boldsymbol{y} + (B(:,\mathcal{I}))^\top \boldsymbol{z} = \hat{\boldsymbol{r}}(\mathcal{I}) \\ & \boldsymbol{y} \geq 0, \boldsymbol{z} \geq -\infty, \end{aligned} \tag{68}$$

which implies that

$$\text{Dual}'_{\mathcal{I}} \leq \text{Dual}_{\mathcal{I}}.$$

On the other hand, any feasible solution to $\text{Dual}'_{\mathcal{I}}$ must be a feasible solution to $\text{Dual}_{\mathcal{I}}$, and we have

$$\text{Dual}'_{\mathcal{I}} \geq \text{Dual}_{\mathcal{I}}.$$

Therefore, we must have $\text{Dual}'_{\mathcal{I}} = \text{Dual}_{\mathcal{I}}$ and our proof is completed. □

E Proof of Theorem 5.2

We first prove the following lemma.

Lemma E.1 *For the optimal basis identified in Algorithm 1 and the corresponding optimal solution $\boldsymbol{q}^*$, we denote by $\mathcal{I}^*$ and $\mathcal{J}^*$ the output sets as long as N_0 satisfies the condition (25). We also denote by $(\boldsymbol{y}^*, \boldsymbol{c}^*)$ the corresponding optimal dual solution. Then, it holds that*

$$N \cdot V_r(\pi^*, \mu_1) - \sum_{n=1}^{N} \hat{\boldsymbol{r}}^\top \mathbb{E}[\boldsymbol{q}^n] \leq \sum_{j \in \mathcal{J}^*} y_j^* \cdot \mathbb{E}[\alpha_j^N] + \sum_{s \in \mathcal{S}} z_s^* \cdot \mathbb{E}[\mu_s^N]. \tag{69}$$

Therefore, it suffices to analyze how the "remaining resources" $(\boldsymbol{\alpha}_{\mathcal{J}^*}^n, \boldsymbol{\mu}^n)$ behave. We now define

$$\tilde{\alpha}_k(n) = \frac{\alpha_k^n}{N-n}, \ \forall k \in \mathcal{J}^* \text{ and } \tilde{\mu}_s(n) = \frac{\mu_s^n}{N-n}, \ \forall s \in \mathcal{S}, \forall n \in [N]. \tag{70}$$

The key is to show that the stochastic process $\tilde{\alpha}_k(n)$ and $\tilde{\mu}_s$ possess some concentration properties such that they will stay within a small neighborhood of their initial value α_k^1 and μ_s^1 for a sufficiently long time. We denote by τ the time that one of $\tilde{\alpha}_k(n)$ for each $k \in \mathcal{J}^*$ and $\tilde{\mu}_s$ for each $s \in \mathcal{S}$ escape this neighborhood. Then, both $\text{Regret}_r(\pi, N)$ and $\text{Regret}_k(\pi, N)$ for each $k \in [K]$ can be upper bounded by $\mathbb{E}[N - \tau]$. From the update rule (22) and (23), we know that

$$\tilde{\alpha}_k(n+1) = \tilde{\alpha}_k(n) - \frac{\sum_{(s,a) \in \mathcal{I}^*} \boldsymbol{c}^n(s,a) \cdot q^n(s,a) - \tilde{\alpha}_k(n)}{N-n-1}, \ \forall k \in \mathcal{J}^* \tag{71}$$

and

$$\tilde{\mu}_s(n+1) = \tilde{\mu}_s(n) - \frac{\sum_{(s,a) \in \mathcal{I}^*} q^n(s,a) \cdot (\delta_{s',s} - \gamma \mathbb{1}_{\{s'=s^n(s,a)\}}) - \tilde{\mu}_s(n)}{N-n-1}. \tag{72}$$

Ideally, both $\tilde{\alpha}_k(n+1)$ and $\tilde{\mu}_s(n+1)$ will have the same expectation as $\tilde{\alpha}_k(n)$ and $\tilde{\mu}_s(n)$ such that they become a martingale. However, this is not true since we have estimation error over $C(\mathcal{J}^*, \mathcal{I}^*)$ and $B(:, \mathcal{I}^*)$, and we only use their estimates to compute $\boldsymbol{q}^n$. Nevertheless, we can show that $\tilde{\alpha}_k(n)$ for each $k \in \mathcal{J}^*$ and $\tilde{\mu}_s$ for each $s \in \mathcal{S}$ behave as a sub-martingale. Then, from the concentration property of the sub-martingale, we upper bound $\mathbb{E}[\alpha_k^N]$ for each $k \in \mathcal{J}^*$ and $\mathbb{E}[\mu_s^N]$ for each $s \in \mathcal{S}$. The term $|\mathbb{E}[\alpha_k^N]|$ for each $k \in [K] \backslash \mathcal{J}^*$ can be upper bounded as well. The results are presented in the following lemma.

Lemma E.2 *Denote by $\bar{\pi}^N$ the output policy of Algorithm 2 and denote by N the number of rounds. Then, it holds that*

$$N \cdot V_r(\pi^*, \mu_1) - \sum_{n=1}^{N} \hat{\boldsymbol{r}}^\top \mathbb{E}[\boldsymbol{q}^n] \leq O\left(\frac{(|\mathcal{S}| + K)^3}{\alpha \cdot \sigma \cdot \min\{\sigma^2, (1-\gamma)^2 \cdot \Delta\}} \cdot \frac{\log(N)}{N} \right)$$

where the parameters $\alpha = \min_{k \in [K]}\{\alpha_k\}$, $\Delta = \min\{\Delta_1^2, \Delta_2^2\}$ with Δ_1 given in (16) and Δ_2 given in (19), σ given in (27). Also, for any $k \in [K]$, we have

$$N \cdot \alpha_k - \sum_{n=1}^{N} \hat{\boldsymbol{c}}_k^\top \mathbb{E}\left[\boldsymbol{q}^n\right] \leq O\left(\frac{(|\mathcal{S}| + K)^3}{\alpha \cdot \sigma \cdot \min\{\sigma^2, (1-\gamma)^2 \cdot \Delta\}} \cdot \frac{\log(N)}{N} \right). \tag{73}$$

E.1 Proof of Lemma E.1

Note that the distribution of $\boldsymbol{q}^n$ is independent of the distribution of $\boldsymbol{r}^n$. We know that

$$\mathbb{E}\left[\sum_{n=1}^{N} (\boldsymbol{r}^n)^\top \boldsymbol{q}^n \right] = \sum_{n=1}^{N} (\hat{\boldsymbol{r}})^\top \mathbb{E}[\boldsymbol{q}^n] = \sum_{n=1}^{N} (\hat{\boldsymbol{r}}_{\mathcal{I}^*})^\top \mathbb{E}\left[\boldsymbol{q}_{\mathcal{I}^*}^n\right]$$

Denote by $\boldsymbol{q}^*$ and $\boldsymbol{y}^*, \boldsymbol{z}^*$ the optimal primal-dual variable corresponding to the optimal basis $\mathcal{I}^*$ and $\mathcal{J}^*$. From the complementary slackness condition and noting that $\boldsymbol{q}_{\mathcal{I}^*}^* > 0$, we know that

$$(C(\mathcal{J}^*, \mathcal{I}^*))^\top \boldsymbol{y}_{\mathcal{J}^*}^* + (B(:, \mathcal{I}^*))^\top \boldsymbol{z}^* = \hat{\boldsymbol{r}}_{\mathcal{I}^*}. \tag{74}$$

Also, we can define a matrix $C^n(\mathcal{J}^*, \mathcal{I}^*)$ such that the element of $C^n(\mathcal{J}^*, \mathcal{I}^*)$ at the $k \in \mathcal{J}^*$ row and $(s,a) \in \mathcal{I}^*$ column is $c_k^n(s,a)$. We can also define a matrix $B^n(:, \mathcal{I}^*)$ such that the element of $B^n(:, \mathcal{I}^*)$ at the $s' \in \mathcal{S}$ row and $(s,a) \in \mathcal{I}^*$ column is $\delta_{s',s} - \gamma \cdot \mathbb{1}_{\{s'=s^n(s,a)\}}$. It is easy to see that

$$\mathbb{E}[C^n(\mathcal{J}^*, \mathcal{I}^*)] = C(\mathcal{J}^*, \mathcal{I}^*) \text{ and } \mathbb{E}[B^n(:, \mathcal{I}^*)] = B(:, \mathcal{I}^*).$$

Then, it holds that

$$\begin{aligned}
\sum_{n=1}^{N} (\hat{\boldsymbol{r}}_{\mathcal{I}^*})^\top \mathbb{E}\left[\boldsymbol{q}_{\mathcal{I}^*}^n\right] &= \sum_{n=1}^{N} \left((C(\mathcal{J}^*, \mathcal{I}^*))^\top \boldsymbol{y}_{\mathcal{J}^*}^* + (B(:, \mathcal{I}^*))^\top \boldsymbol{z}^* \right)^\top \mathbb{E}[\boldsymbol{q}_{\mathcal{I}^*}^n] \\
&= \mathbb{E}\left[\sum_{n=1}^{N} \left((C^n(\mathcal{J}^*, \mathcal{I}^*))^\top \boldsymbol{y}_{\mathcal{J}^*}^* + (B^n(:, \mathcal{I}^*))^\top \boldsymbol{z}^* \right)^\top \boldsymbol{q}_{\mathcal{I}^*}^n \right] \\
&= \mathbb{E}\left[\sum_{n=1}^{N} \left((\boldsymbol{y}^*)^\top (C^n(\mathcal{J}^*, \mathcal{I}^*)) \boldsymbol{q}_{\mathcal{I}^*}^n + (\boldsymbol{z}^*)^\top (B^n(:, \mathcal{I}^*)) \boldsymbol{q}_{\mathcal{I}^*}^n \right) \right]
\end{aligned} \tag{75}$$

Note that we have

$$\sum_{n=1}^{N} C^n(\mathcal{J}^*, \mathcal{I}^*)) \boldsymbol{q}_{\mathcal{I}^*}^n = \boldsymbol{\alpha}_{\mathcal{J}^*}^1 - \boldsymbol{\alpha}^N \tag{76}$$

and

$$\sum_{n=1}^{N} B^n(:, \mathcal{I}^*)) \boldsymbol{q}_{\mathcal{I}^*}^n = \boldsymbol{\mu}^1 - \boldsymbol{\mu}^N. \tag{77}$$

Plugging (76) and (77) back into (75), we get that

$$\sum_{n=1}^{N} (\hat{\boldsymbol{r}}_{\mathcal{I}^*})^\top \mathbb{E}\left[\boldsymbol{q}_{\mathcal{I}^*}^n\right] = (\boldsymbol{y}_{\mathcal{J}^*}^*)^\top \boldsymbol{\alpha}_{\mathcal{J}^*}^1 + (\boldsymbol{z}^*)^\top \boldsymbol{\mu}^1 - (\boldsymbol{y}_{\mathcal{J}^*}^*)^\top \mathbb{E}\left[\boldsymbol{\alpha}_{\mathcal{J}^*}^N\right] - (\boldsymbol{z}^*) \top \mathbb{E}\left[\boldsymbol{\mu}^N\right].$$

Note that

$$V_r(\pi^*, \mu_1) = (\boldsymbol{y}_{\mathcal{J}^*}^*)^\top \boldsymbol{\alpha}_{\mathcal{J}^*}^1 + (\boldsymbol{z}^*)^\top \boldsymbol{\mu}^1$$

that holds from the strong duality of V^{Infi} (4). Then, we have that

$$N \cdot V_r(\pi^*, \mu_1) - \sum_{n=1}^{N} (\hat{\boldsymbol{r}})^\top \mathbb{E}[\boldsymbol{q}^n] \leq (\boldsymbol{y}_{\mathcal{J}^*}^*)^\top \mathbb{E}\left[\boldsymbol{\alpha}_{\mathcal{J}^*}^N\right] + (\boldsymbol{z}^*)^\top \mathbb{E}\left[\boldsymbol{\mu}^N\right]. \tag{78}$$

Our proof is thus completed.

E.2 Proof of Lemma E.2

We now condition on the event that $|\bar{r}_n(s,a) - \hat{r}(s,a)|$, $|\bar{c}_{k,n}(s,a) - \hat{c}_k(s,a)|$, and $|\bar{P}_n(s'|s,a) - P(s'|s,a)|$ for each $k \in [K]$, $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $s' \in \mathcal{S}$ are all bounded by $\text{Rad}(n, \varepsilon)$, for any $n \in [N]$. We know that this event happens with probability at least $1 - N \cdot (K|\mathcal{S}||\mathcal{A}| - |\mathcal{S}|^2|\mathcal{A}|) \cdot \varepsilon$. From now on, we set $\varepsilon = \frac{1}{N^2}$ to guarantee that the event happens with probability at least $1 - \frac{K|\mathcal{S}||\mathcal{A}| + |\mathcal{S}|^2|\mathcal{A}|}{N}$.

We consider the stochastic process $\tilde{\alpha}_k(n)$ and $\tilde{\mu}_s(n)$ defined in (70). For a fixed $\nu > 0$ which we specify later, we define a set

$$\mathcal{X} = \{\boldsymbol{\alpha}' \in \mathbb{R}^{|\mathcal{J}^*|} : \alpha_k' \in [\alpha_k - \nu, \alpha_k + \nu], \forall k \in \mathcal{J}^*\}, \tag{79}$$

and

$$\mathcal{Y} = \{\boldsymbol{\mu}' \in \mathbb{R}^{|\mathcal{S}|} : \mu_s' \in [\mu_s - \nu, \mu_s + \nu], \forall s \in \mathcal{S}\}. \tag{80}$$

It is easy to see that initially, $\tilde{\boldsymbol{\alpha}}(1) \in \mathcal{X}$ and $\tilde{\boldsymbol{\mu}}(1) \in \mathcal{Y}$. We show that $\tilde{\boldsymbol{\alpha}}(n)$ and $\tilde{\boldsymbol{\mu}}(n)$ behave well as long as they stay in the region $\mathcal{X}$ and $\mathcal{Y}$ for a sufficiently long time. To this end, we define a stopping time

$$\tau = \min_{n \in [N]} \{\tilde{\boldsymbol{\alpha}}(n) \notin \mathcal{X} \text{ or } \tilde{\boldsymbol{\mu}}(n) \notin \mathcal{Y}\}. \tag{81}$$

Note that in Algorithm 2, to stop $\boldsymbol{q}^n$ from behaving ill when n is small, we project it to a set that guarantees $\|\boldsymbol{q}^n\|_1 \leq 2$. We now show in the following claim that when n is large enough but smaller than the stopping time τ, there is no need to do projection.

Claim E.3 *There exist two constants N_0' and ν_0. When $\max\{N_0, N_0'\} \leq n \leq \tau$, and $\nu \leq \nu_0$, it holds that $\|\tilde{\boldsymbol{q}}_{\mathcal{I}^*}^n\|_1 \leq 2$, where $\tilde{\boldsymbol{q}}_{\mathcal{I}^*}^n$ denotes the solution to the linear equations (21). Specifically, N_0 is given in (25) and N_0' is given as follows*

$$N_0' = 16 \cdot \frac{(|\mathcal{S}| + K)^2}{\sigma^2} \cdot \log(1/\varepsilon) \tag{82}$$

Also, ν_0 is given as follows

$$\nu_0 := 16 \cdot \frac{\sigma \cdot (1-\gamma)}{(|\mathcal{S}| + K)^2}. \tag{83}$$

We set ν satisfy the condition $\nu \leq \nu_0$ with ν_0 satisfies the condition in Claim E.3. We bound $\mathbb{E}[N - \tau]$ in the following claim.

Claim E.4 *Let the stopping time τ be defined in (81). It holds that*

$$\mathbb{E}[N - \tau] \leq \max\{N_0, N_0'\} + 2(K + |\mathcal{S}|) \cdot \exp(-\nu^2/8)$$

where N_0 is given in (25) and N_0' is given in (82), as long as

$$N \geq \max\{N_0, N_0'\} \;\; \textit{and} \;\; N \geq \frac{8}{\nu^2} \geq \frac{8}{\nu_0^2} = \frac{(|\mathcal{S}| + K)^4}{32\sigma^2 \cdot (1-\gamma)^2}. \tag{84}$$

Also, for any N' such that $\max\{N_0, N_0'\} \leq N' \leq N$, it holds that

$$P(\tau \leq N') \leq \frac{(K + |\mathcal{S}|) \cdot \nu^2}{4} \cdot \exp\left(-\frac{\nu^2 \cdot (N - N' + 1)}{8}\right). \tag{85}$$

From the definition of the stopping time τ in (81), we know that for each $k \in \mathcal{J}^*$, it holds

$$\alpha_k^{\tau-1} \in [(N - \tau + 1) \cdot (\alpha_k - \nu), (N - \tau + 1) \cdot (\alpha_k + \nu)]$$

Thus, we have that

$$|\alpha_k^N| \leq |\alpha_k^{\tau-1}| + \sum_{t=\tau}^{N} \sum_{(s,a) \in \mathcal{I}^*} c_k^n(s,a) \cdot q^n(s,a) \tag{86}$$

and thus

$$\left|\mathbb{E}[\alpha_k^N]\right| \leq 4\mathbb{E}[N - \tau] \leq 4\max\{N_0, N_0'\} + 8(K + |\mathcal{S}|) \cdot \exp(-\nu^2/8). \tag{87}$$

Following the same procedure, we can show that for each $s \in \mathcal{S}$, it holds that

$$\left|\mathbb{E}[\mu^N(s)]\right| \leq 4\mathbb{E}[N - \tau] \leq 4\max\{N_0, N_0'\} + 8(K + |\mathcal{S}|) \cdot \exp(-\nu^2/8). \tag{88}$$

We finally consider the other constraints $k \in \mathcal{J}^{*c}$. Note that following the definition of $\boldsymbol{\alpha}^n$ and $\boldsymbol{\mu}^n$, we have that

$$A^* \cdot \left(\sum_{n=1}^{N} \mathbb{E}[\boldsymbol{q}_{\mathcal{I}^*}^n]\right) = [\boldsymbol{\alpha}_{\mathcal{J}^*}^1 - \mathbb{E}\left[\boldsymbol{\alpha}_{\mathcal{J}^*}^N\right]; \boldsymbol{\mu}^1 - \mathbb{E}\left[\boldsymbol{\mu}^N\right]]. \tag{89}$$

Also, from the bindingness of $\boldsymbol{q}^*$ regarding the optimal basis $\mathcal{I}^*$ and $\mathcal{J}^*$, we have

$$N \cdot A^* \cdot \boldsymbol{q}_{\mathcal{I}^*}^* = [\boldsymbol{\alpha}_{\mathcal{J}^*}^1; \boldsymbol{\mu}^1]. \tag{90}$$

Therefore, it holds that

$$\sum_{n=1}^{N} \mathbb{E}\left[\boldsymbol{q}_{\mathcal{I}^*}^n\right] = N \cdot \boldsymbol{q}_{\mathcal{I}^*}^* - (A^*)^{-1} \cdot \left[\mathbb{E}\left[\boldsymbol{\alpha}_{\mathcal{J}^*}^N\right]; \mathbb{E}\left[\boldsymbol{\mu}^N\right]\right], \tag{91}$$

and

$$\left\|\sum_{n=1}^{N} \mathbb{E}\left[\boldsymbol{q}_{\mathcal{I}^{*c}}^n\right]\right\|_1 \leq 2N_0. \tag{92}$$

Finally, for any $k \in \mathcal{J}^{*c}$, we have

$$\begin{aligned}(\hat{\boldsymbol{c}}_k)^\top \left(\sum_{n=1}^{N} \mathbb{E}\left[\boldsymbol{q}^n\right]\right) &= N \cdot (\hat{\boldsymbol{c}}_k)^\top \boldsymbol{q}_{\mathcal{I}^*}^* - (\hat{\boldsymbol{c}}_k)^\top \cdot (A^*)^{-1} \cdot \left[\mathbb{E}\left[\boldsymbol{\alpha}_{\mathcal{J}^*}^N\right]; \mathbb{E}\left[\boldsymbol{\mu}^N\right]\right] + 2N_0 \\ &= N \cdot (\hat{\boldsymbol{c}}_k)^\top \boldsymbol{q}^* - (\hat{\boldsymbol{c}}_k)^\top \cdot (A^*)^{-1} \cdot \left[\mathbb{E}\left[\boldsymbol{\alpha}_{\mathcal{J}^*}^N\right]; \mathbb{E}\left[\boldsymbol{\mu}^N\right]\right] + 2N_0\end{aligned} \tag{93}$$

From the feasibility of $\boldsymbol{q}^*$, we know that

$$N \cdot \alpha_k \geq N \cdot (\hat{\boldsymbol{c}}_k)^\top \boldsymbol{q}^*.$$

Therefore, for any $k \in \mathcal{J}^{*c}$, it holds that

$$\begin{aligned} N \cdot \alpha_k - \sum_{n=1}^{N} \hat{\boldsymbol{c}}_k^\top \mathbb{E}\left[\boldsymbol{q}^n\right] &\leq (\hat{\boldsymbol{c}}_k)^\top \cdot (A^*)^{-1} \cdot \left[\mathbb{E}[\boldsymbol{\alpha}_{\mathcal{J}^*}^N]; \mathbb{E}[\boldsymbol{\mu}^N]\right] + 2N_0 \\ &\leq \frac{K + |\mathcal{S}|}{\sigma} \cdot \left(4\max\{N_0, N_0'\} + 8(K + |\mathcal{S}|) \cdot \exp(-\nu^2/8)\right) + 2N_0. \end{aligned} \tag{94}$$

Moreover, the definition of σ in (27) implies that following upper bound on the norm of the dual variable $\boldsymbol{y}^*$ and $\boldsymbol{z}^*$.

$$\|\boldsymbol{y}^*\|_1 + \|\boldsymbol{z}^*\|_1 = \|(A^{*\top})^{-1} \cdot \hat{\boldsymbol{r}}_{\mathcal{J}^*}\|_1 \leq \frac{|\mathcal{S}| + K}{\sigma}. \tag{95}$$

Therefore, we know that the regret over the reward and the regret over the constraint violation can all be bounded by using (87), (88), and (94). We present the bounds as follows.

$$N \cdot V_r(\pi^*, \mu_1) - \sum_{n=1}^{N} \hat{\boldsymbol{r}}^\top \mathbb{E}[\boldsymbol{q}^n] \leq \frac{K + |\mathcal{S}|}{\sigma} \cdot \left(4\max\{N_0, N_0'\} + 8(K + |\mathcal{S}|) \cdot \exp(-\nu^2/8)\right). \tag{96}$$

Meanwhile, the constraint violations are bounded by (87), (88), and (94). Our proof is thus completed.

Proof of Claim E.3. Denote by $\boldsymbol{q}^*$ the optimal solution corresponding to the optimal basis $\mathcal{I}^*$ and $\mathcal{J}^*$. Then, it holds that

$$\begin{bmatrix} C(\mathcal{J}^*, \mathcal{I}^*) \\ B(:, \mathcal{I}^*) \end{bmatrix} \cdot \boldsymbol{q}^*_{\mathcal{I}^*} = \begin{bmatrix} \boldsymbol{\alpha}_{\mathcal{J}^*} \\ \boldsymbol{\mu} \end{bmatrix}. \tag{97}$$

We compare $\tilde{\boldsymbol{q}}^n_{\mathcal{I}^*}$ with $\boldsymbol{q}^*_{\mathcal{I}^*}$ when n large enough. Note that when $n \geq N_0$, $\tilde{\boldsymbol{q}}^n$ is the solution to the following linear equations

$$\begin{bmatrix} \bar{C}^n(\mathcal{J}^*, \mathcal{I}^*) \\ \bar{B}^n(:, \mathcal{I}^*) \end{bmatrix} \cdot \tilde{\boldsymbol{q}}^n_{\mathcal{I}^*} = \begin{bmatrix} \dfrac{\boldsymbol{\alpha}^n_{\mathcal{J}^*}}{N - n + 1} \\ \dfrac{\boldsymbol{\mu}^n}{N - n + 1} \end{bmatrix}. \tag{98}$$

When $n \leq \tau$, we know that

$$\left| \boldsymbol{\alpha}_{\mathcal{J}^*} - \frac{\boldsymbol{\alpha}^n_{\mathcal{J}^*}}{N - n + 1} \right| \leq \nu \tag{99}$$

and

$$\left| \boldsymbol{\mu} - \frac{\boldsymbol{\mu}^n}{N - n + 1} \right| \leq \nu. \tag{100}$$

Moreover, we know that the absolute value of each element of $\bar{C}^n(\mathcal{J}^*, \mathcal{I}^*) - C(\mathcal{J}^*, \mathcal{I}^*)$, and $\bar{B}^n(:, \mathcal{I}^*) - B(:, \mathcal{I}^*)$ is upper bounded by $\mathrm{Rad}(n, \varepsilon)$. We now bound the distance between the solutions to the linear equations (97) and (98). The perturbation of the matrix is denoted as

$$\Delta A^* = \begin{bmatrix} C(\mathcal{J}^*, \mathcal{I}^*) - \bar{C}^n(\mathcal{J}^*, \mathcal{I}^*) \\ B(:, \mathcal{I}^*) - \bar{B}(:, \mathcal{I}^*) \end{bmatrix}.$$

Clearly, it holds that

$$\|\Delta A^*\|_1 \leq \mathrm{Rad}(n, \varepsilon) \cdot (K + |\mathcal{S}|). \tag{101}$$

Therefore, as long as

$$\|\Delta A^*\|_1 \leq \mathrm{Rad}(n, \varepsilon) \cdot (K + |\mathcal{S}|) \leq \frac{1}{2\|(A^*)^{-1}\|_1} \leq \frac{1}{2\sigma}, \tag{102}$$

following standard perturbation analysis of linear equations [28], we have that

$$\begin{aligned}\frac{\|\tilde{\boldsymbol{q}}^n_{\mathcal{I}^*} - \boldsymbol{q}^*_{\mathcal{I}^*}\|_1}{\|\boldsymbol{q}^*_{\mathcal{I}^*}\|_1} &\leq \frac{\kappa(A^*)}{1-\kappa(A^*)\cdot\frac{\|\Delta A^*\|_1}{\|A^*\|_1}}\cdot\left(\frac{\|\Delta A^*\|_1}{\|A^*\|_1}+\frac{(|\mathcal{S}+K|\cdot\nu)}{\|[\boldsymbol{\alpha}_{\mathcal{J}^*};\boldsymbol{\mu}]\|_1}\right)\\ &\leq 2\cdot\kappa(A^*)\cdot\left(\frac{\|\Delta A^*\|_1}{\|A^*\|_1}+\frac{(|\mathcal{S}+K|\cdot\nu)}{\|[\boldsymbol{\alpha}_{\mathcal{J}^*};\boldsymbol{\mu}]\|_1}\right) \qquad (103)\\ &\leq 2\cdot\kappa(A^*)\cdot\left(\frac{\|\Delta A^*\|_1}{\|A^*\|_1}+\frac{(|\mathcal{S}|+K))\cdot\nu}{1-\gamma}\right),\end{aligned}$$

where $\kappa(A^*) = \|A^*\|_1\cdot\|(A^*)^{-1}\|_1$ denotes the conditional number of A^*. The last inequality follows from $\|[\boldsymbol{\alpha}_{\mathcal{J}^*};\boldsymbol{\mu}]\|_1 \geq 1-\gamma$. Further, note that $\|\boldsymbol{q}^*_{\mathcal{I}^*}\|_1 = 1$. Therefore, in order to satisfy the condition $\|\tilde{\boldsymbol{q}}^n_{\mathcal{I}^*}\|_1 \leq 2$, we only need the right hand side of (103) to be upper bounded by 1. Clearly, as long as n satisfies the condition (102) and the following condition

$$2\cdot\kappa(A^*)\cdot\frac{\|\Delta A^*\|_1}{\|A\|_1} \leq 2\cdot\frac{\mathrm{Rad}(n,\varepsilon)\cdot(K+|\mathcal{S}|)}{\sigma} \leq \frac{1}{2}, \qquad (104)$$

we only need to select a ν such that

$$2\cdot\kappa(A^*)\cdot\frac{(|\mathcal{S}|+K)\cdot\nu}{1-\gamma} \leq \frac{1}{2}. \qquad (105)$$

Combining (102) and (104), we know that n needs to satisfy the following conditions: $n \geq N_0$ and

$$n \geq N_0' := 16\cdot\frac{(|\mathcal{S}|+K)^2}{\sigma^2}\cdot\log(1/\varepsilon). \qquad (106)$$

Also, ν is selected to satisfy the following condition

$$\nu \leq \nu_0 := 16\cdot\frac{\sigma\cdot(1-\gamma)}{(|\mathcal{S}|+K)^2}. \qquad (107)$$

Our proof is thus completed. □

Proof of Claim E.4. Now we fix a $k \in \mathcal{J}^*$. We specify a $\bar{N}_0 = \max\{N_0, N_0'\}$. For any $\bar{N}_0 \leq N' \leq N$, it holds that

$$\tilde{\alpha}_k(N') - \tilde{\alpha}_k(\bar{N}_0) = \sum_{n=\bar{N}_0}^{N'-1}(\tilde{\alpha}_k(n+1) - \tilde{\alpha}_k(n)).$$

We define $\xi_k(n) = \tilde{\alpha}_k(n+1) - \tilde{\alpha}_k(n)$. Then, we have

$$\tilde{\alpha}_k(N') - \tilde{\alpha}_k(\bar{N}_0) = \sum_{n=\bar{N}_0}^{N'-1}(\xi_k(n) - \mathbb{E}[\xi_k(n)|\mathcal{F}_n]) + \sum_{n=\bar{N}_0}^{N'-1}\mathbb{E}[\xi_k(n)|\mathcal{F}_n].$$

where $\mathcal{F}_n$ denotes the filtration of information up to step n. Note that due to the update in (71), we have

$$\xi_k(n) = \frac{\tilde{\alpha}_k(n) - \sum_{(s,a)\in\mathcal{I}^*} c^n_k(s,a)\cdot q^n(s,a)}{N-n-1}.$$

Then, it holds that

$$|\xi_k(n) - \mathbb{E}[\xi_k(n)|\mathcal{F}_n]| \leq \frac{2}{N-n+1} \qquad (108)$$

where the inequality follows from the fact that the value of $\tilde{\alpha}_k(n)$ is deterministic given the filtration $\mathcal{F}_n$ and $\|\boldsymbol{q}^n\|_1 \leq 2$ for any n. Note that

$$\{\xi_k(n) - \mathbb{E}[\xi_k(n)|\mathcal{F}_n]\}_{\forall n=\bar{N}_0,\ldots,N'}$$

forms a martingale difference sequence. Following Hoeffding's inequality, for any $N'' \leq N'$ and any $b > 0$, it holds that

$$\begin{aligned}P\left(\left|\sum_{n=\bar{N}_0}^{N''}(\xi_k(n) - \mathbb{E}[\xi_k(n)|\mathcal{F}_n])\right| \geq b\right) &\leq 2\exp\left(-\frac{b^2}{2\cdot\sum_{n=\bar{N}_0}^{N''}1/(N-n+1)^2}\right)\\ &\leq 2\exp\left(-\frac{b^2\cdot(N-N''+1)}{2}\right).\end{aligned}$$

Therefore, we have that

$$P\left(\left|\sum_{n=\bar{N}_0}^{N''}(\xi_k(n)-\mathbb{E}[\xi_k(n)|\mathcal{F}_n])\right|\geq b \text{ for some } \bar{N}_0\leq N''\leq N'\right)$$
$$\leq \sum_{N''=\bar{N}_0}^{N'} 2\exp\left(-\frac{b^2\cdot(N-N''+1)}{2}\right)\leq b^2\cdot\exp\left(-\frac{b^2\cdot(N-N'+1)}{2}\right) \tag{109}$$

holds for any $b>0$.

We now bound the probability that $\tau>N'$ for one particular N' such that $\bar{N}_0\leq N'\leq N$. Suppose that $N'\leq\tau$, then, from Claim E.3, for each $n\leq N'$, we know that $\|\tilde{\boldsymbol{q}}^n\|_1\leq 2$ and therefore $\boldsymbol{q}^n=\tilde{\boldsymbol{q}}^n$ as the solution to (21). We have

$$\tilde{\alpha}_k(n)=\sum_{(s,a)\in\mathcal{I}^*}\bar{c}_{k,n}(s,a)\cdot q^n(s,a).$$

It holds that

$$|\mathbb{E}[\xi_k(n)|\mathcal{F}_n]|\leq\frac{1}{N-n+1}\cdot\sum_{(s,a)\in\mathcal{I}^*}q^n(s,a)\cdot|\mathbb{E}[\bar{c}_{k,n}(s,a)]-\hat{c}_k^n(s,a)|\leq\frac{2\mathrm{Rad}(n,\varepsilon)}{N-n+1}. \tag{110}$$

Then, we know that

$$\begin{aligned}\sum_{n=\bar{N}_0}^{N'-1}|\mathbb{E}[\xi_k(n)|\mathcal{F}_n]|&\leq\sqrt{\frac{\log(2/\varepsilon)}{2}}\cdot\sum_{n=\bar{N}_0}^{N'-1}\frac{1}{\sqrt{n}\cdot(N-n)}\leq\sqrt{\frac{\log(2/\varepsilon)}{2}}\cdot\sqrt{N'-1}\cdot\sum_{n=\bar{N}_0}^{N'-1}\frac{1}{n\cdot(N-n)}\\&=\sqrt{\frac{\log(2/\varepsilon)}{2}}\cdot\frac{\sqrt{N'-1}}{N}\cdot\sum_{n=\bar{N}_0}^{N'-1}\left(\frac{1}{n}+\frac{1}{N-n}\right)\\&\leq\sqrt{2\log(2/\varepsilon)}\cdot\frac{\sqrt{N'-1}}{N}\cdot\log(N)\leq\frac{\sqrt{2\log(2/\varepsilon)}}{\sqrt{N}}\cdot\log(N)\leq\frac{\nu}{2}\end{aligned} \tag{111}$$

for a N large enough such that

$$N\geq\frac{8}{\nu^2}\geq\frac{8}{\nu_0^2}=\frac{(|\mathcal{S}|+K)^4}{32\sigma^2\cdot(1-\gamma)^2} \tag{112}$$

Combining (111) and (109) with $b=\nu/2$, and apply a union bound over all $k\in\mathcal{J}^*$ and $s\in\mathcal{S}$, we know that

$$P(\tau\leq N')\leq\frac{(K+|\mathcal{S}|)\cdot\nu^2}{4}\cdot\exp\left(-\frac{\nu^2\cdot(N-N'+1)}{8}\right). \tag{113}$$

Therefore, we know that

$$\mathbb{E}[N-\tau]=\sum_{N'=1}^{N}P(\tau\leq N')\leq\bar{N}_0+\sum_{N'=\bar{N}_0}^{N}P(\tau\leq N')\leq\bar{N}_0+2(K+|\mathcal{S}|)\cdot\exp(-\nu^2/8)$$

which completes our proof. □

E.3 Final Proof of Theorem 5.2

Note that in the proof of Lemma E.2, we have shown the following bounds.

$$\left|\mathbb{E}[\alpha_k^N]\right|\leq 4\mathbb{E}[N-\tau]\leq 4\max\{N_0,N_0'\}+8(K+|\mathcal{S}|)\cdot\exp(-\nu^2/8). \tag{114}$$

holds for the each $k\in\mathcal{J}^*$. For each $s\in\mathcal{S}$, it holds that

$$\left|\mathbb{E}[\mu^N(s)]\right|\leq 4\mathbb{E}[N-\tau]\leq 4\max\{N_0,N_0'\}+8(K+|\mathcal{S}|)\cdot\exp(-\nu^2/8). \tag{115}$$

The caveat of directly transferring the bound of (114) and (115) into the sample complexity bounds of the policy $\bar{\pi}^N$ is that, the vector $\bar{\boldsymbol{q}}^N$ does not directly characterize an occupancy measure. This point

can be seen by noting that there is a gap between $B\bar{\boldsymbol{q}}^N$ and $\boldsymbol{\mu}$, though bounded by $O(\log(N)/N)$ by setting $\varepsilon = 1/N^2$. However, we can show that the gap between $\bar{\boldsymbol{q}}^N$ and $\boldsymbol{q}^*$ is upper bounded by $O(\log(N)/N)$, which implies a bound over the gap between the policy $\bar{\pi}^N$ and the optimal policy π^* that corresponds to the occupancy measure $\boldsymbol{q}^*$. This bound over the gap between the policy distributions can be then transferred into the bound over the gap between the state-value functions under the policy $\bar{\pi}^N$ and π^*. The regret bounds can be obtained then.

We first bound the gap between $\bar{\boldsymbol{q}}^N$ and $\boldsymbol{q}^*$. Note that as long as $n \geq N_0$, we have $\mathcal{I}_n = \mathcal{I}^*$ following Theorem 5.1. Then, by noting $\mathcal{I}_N = \mathcal{I}^*$, we know that

$$\bar{\boldsymbol{q}}^N_{\mathcal{I}^{*c}} = \boldsymbol{q}^*_{\mathcal{I}^{*c}}. \tag{116}$$

Also, note that following the definition of $\boldsymbol{\alpha}^n$ and $\boldsymbol{\mu}^n$, we have that

$$A^* \cdot \left(\sum_{n=1}^{N} \mathbb{E}[\boldsymbol{q}^n_{\mathcal{I}^*}] \right) = [\boldsymbol{\alpha}^1_{\mathcal{J}^*} - \mathbb{E}\left[\boldsymbol{\alpha}^N_{\mathcal{J}^*}\right]; \boldsymbol{\mu}^1 - \mathbb{E}\left[\boldsymbol{\mu}^N\right]]. \tag{117}$$

Also, from the bindingness of $\boldsymbol{q}^*$ regarding the optimal basis $\mathcal{I}^*$ and $\mathcal{J}^*$, we have

$$N \cdot A^* \cdot \boldsymbol{q}^*_{\mathcal{I}^*} = [\boldsymbol{\alpha}^1_{\mathcal{J}^*}; \boldsymbol{\mu}^1]. \tag{118}$$

Then, from (117) and (118), we know that

$$\begin{aligned} \left\| \mathbb{E}\left[\bar{\boldsymbol{q}}^N_{\mathcal{I}^*}\right] - \boldsymbol{q}^*_{\mathcal{I}^*} \right\|_\infty &= \left\| \boldsymbol{q}^*_{\mathcal{I}^*} - \frac{1}{N} \cdot \sum_{n=1}^{N} \mathbb{E}\left[\boldsymbol{q}^n_{\mathcal{I}^*}\right] \right\|_\infty = \frac{\left\| (A^*)^{-1} \cdot \left[\mathbb{E}\left[\boldsymbol{\alpha}^N_{\mathcal{J}^*}\right]; \mathbb{E}\left[\boldsymbol{\mu}^N\right]\right] \right\|_\infty}{N} \\ &\leq \frac{\left\| \left[\mathbb{E}\left[\boldsymbol{\alpha}^N_{\mathcal{J}^*}\right]; \mathbb{E}\left[\boldsymbol{\mu}^N\right]\right] \right\|_\infty}{\sigma \cdot N} \\ &\leq \frac{1}{\sigma \cdot N} \cdot \left(4 \max\{N_0, N_0'\} + 8(K + |\mathcal{S}|) \cdot \exp(-\nu^2/8)\right). \end{aligned} \tag{119}$$

From Markov's inequality, for each $i \in \mathcal{I}^*$ and any $a > 0$, we know that

$$P\left(|\bar{\boldsymbol{q}}^N_i - \boldsymbol{q}^*_i| > g\right) \leq \frac{1}{g \cdot \sigma \cdot N} \cdot \left(4 \max\{N_0, N_0'\} + 8(K + |\mathcal{S}|) \cdot \exp(-\nu^2/8)\right). \tag{120}$$

We denote by

$$\xi = \min_{(s,a) \in \mathcal{I}^*} \{q^*(s,a)\}. \tag{121}$$

The policy $\bar{\pi}^N$ is essentially random by noting that $\boldsymbol{q}^N$ is a random variable, where the randomness comes from the randomness of the filtration $\mathcal{F}_N$. For each $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we denote by $\bar{\pi}(a|s)$ the (ex-ante) probability that the random policy $\bar{\pi}$ takes the action a given the state s. Then, for any $0 < g \leq \xi/2$, we note that

$$|\boldsymbol{q}^N_i - \boldsymbol{q}^*_i| \leq g \text{ for each } i = (s,a) \in \mathcal{I}^* \text{ implies that } \left| \frac{\boldsymbol{q}^N(s,a)}{\sum_{a' \in \mathcal{A}} \boldsymbol{q}^N(s,a')} - \frac{\boldsymbol{q}^*(s,a)}{\sum_{a' \in \mathcal{A}} \boldsymbol{q}^*(s,a')} \right| \leq \frac{2g}{\xi}, \tag{122}$$

for each $i = (s,a) \in \mathcal{I}^*$. For any $0 < g \leq \xi/2$, note that

$$P\left(|\boldsymbol{q}^N_i - \boldsymbol{q}^*_i| \leq g \text{ for each } i = (s,a) \in \mathcal{I}^*\right) \geq 1 - \frac{|\mathcal{S}| + K}{g \cdot \sigma \cdot N} \cdot \left(4 \max\{N_0, N_0'\} + 8(K + |\mathcal{S}|) \cdot \exp(-\nu^2/8)\right), \tag{123}$$

where the inequality follows from the bound (120) and the union bound over $i \in \mathcal{I}^*$. Therefore, for any $0 < g \leq \xi/2$ and any (s,a), we know that

$$P\left(\left|\bar{\pi}^N(a|s) - \pi^*(a|s)\right| \leq \frac{2g}{\xi}\right) \geq 1 - \frac{|\mathcal{S}| + K}{g \cdot \sigma \cdot N} \cdot \left(4 \max\{N_0, N_0'\} + 8(K + |\mathcal{S}|) \cdot \exp(-\nu^2/8)\right). \tag{124}$$

From the above inequality, for any (s,a), we have that

$$\begin{aligned}
&\left|\mathbb{E}\left[\bar{\pi}^N(a|s)\right]-\pi^*(a|s)\right| \\
\leq&\mathbb{E}\left[\left|\bar{\pi}^N(a|s)-\pi^*(a|s)\right|\right] \leq \frac{2}{N\xi}+\frac{2}{\xi}\cdot\int_{g=\frac{1}{N}}^{\xi/2} P\left(\left|\bar{\pi}^N(a|s)-\pi^*(a|s)\right|\geq\frac{2g}{\xi}\right)dg \\
\leq&\frac{2}{N\xi}+\frac{2(|\mathcal{S}|+K)}{\xi\cdot\sigma\cdot N}\cdot\left(4\max\{N_0,N_0'\}+8(K+|\mathcal{S}|)\cdot\exp(-\nu^2/8)\right)\cdot\int_{g=\frac{1}{N}}^{\xi/2}\frac{dg}{g} \\
=&\frac{2}{N\xi}+\frac{2(|\mathcal{S}|+K)}{\xi\cdot\sigma\cdot N}\cdot\left(4\max\{N_0,N_0'\}+8(K+|\mathcal{S}|)\cdot\exp(-\nu^2/8)\right)\cdot\left(\log(N)+\log(\xi/2)\right).
\end{aligned} \tag{125}$$

We finally transfer the bound (125) into the sample complexity bounds of policy $\bar{\pi}^N$. We use the state-value functions $V_r(\pi,s)$, defined for any initial state s and any policy π as follows

$$V_r(\pi,s)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\cdot r(s_t,a_t)\mid s\right], \tag{126}$$

where (s_t,a_t) is generated according to the policy π and the transition kernel P with the initial state s. Note that the value of $V_r(\pi,s)$ for any $s\in\mathcal{S}$ can be obtained from solving Bellman's equation under policy π

$$V_r(\pi,s)=\mathbb{E}_{a\sim\pi(\cdot|s)}\left[\hat{r}(s,a)+\gamma\cdot\mathbb{E}_{s'\sim P(\cdot|s,a)}[V_r(\pi,s')]\right]. \tag{127}$$

We define a matrix $B^\pi\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ such that the s-th row s'-th column element is

$$B^\pi(s,s')=\delta_{s,s'}-\gamma\cdot\sum_{a\in\mathcal{A}}\pi(a|s)\cdot P(s'|s,a). \tag{128}$$

Then, the matrix B^π represents the state transition probability matrix under the policy π. Denote by

$$\boldsymbol{V}_r(\pi)=(V_r(\pi,s))_{\forall s\in\mathcal{S}}$$

and

$$\hat{\boldsymbol{r}}(\pi)=(\sum_{a\in\mathcal{A}}\pi(a|s)\cdot\hat{r}(s,a))_{\forall s\in\mathcal{S}}.$$

We have that the state values $\boldsymbol{V}_r(\pi)$ is the solution to the linear equation

$$B^\pi\boldsymbol{V}_r(\pi)=\hat{\boldsymbol{r}}(\pi) \tag{129}$$

To bound the regret, we bound the solution to the linear equation (129) with π being $\bar{\pi}^N$ and π^* separately. The perturbation of the right hand of the equation (129) is

$$\Delta\hat{\boldsymbol{r}}=\hat{\boldsymbol{r}}(\bar{\pi}^N)-\hat{\boldsymbol{r}}(\pi^*).$$

Clearly, we have that

$$\|\Delta\hat{\boldsymbol{r}}\|_\infty\leq\frac{2}{N\xi}+\frac{2(|\mathcal{S}|+K)}{\xi\cdot\sigma\cdot N}\cdot\left(4\max\{N_0,N_0'\}+8(K+|\mathcal{S}|)\cdot\exp(-\nu^2/8)\right)\cdot\left(\log(N)+\log(\xi/2)\right). \tag{130}$$

The perturbation of the matrix is denoted as

$$\Delta B=B^{\bar{\pi}^N}-B^{\pi^*}.$$

Clearly, it holds that

$$\|\Delta B\|_\infty\leq\frac{2\gamma}{N\xi}+\frac{2\gamma(|\mathcal{S}|+K)}{\xi\cdot\sigma\cdot N}\cdot\left(4\max\{N_0,N_0'\}+8(K+|\mathcal{S}|)\cdot\exp(-\nu^2/8)\right)\cdot\left(\log(N)+\log(\xi/2)\right). \tag{131}$$

We plug the formulation of N_0 in (25) and N_0' in (82) into the bound (130) and (131). We obtain

$$\|\Delta\hat{\boldsymbol{r}}\|_\infty\leq C_1\cdot\frac{(|\mathcal{S}|+K)^3}{\alpha^2\cdot\xi\cdot\sigma\cdot\min\{\sigma^2,(1-\gamma)^2\cdot\Delta\}}\cdot\frac{\log^2(N)}{N} \tag{132}$$

where C_1 is a constant, $\alpha = \min_{k\in[K]}\{\alpha_k\}$, and $\Delta = \min\{\Delta_1^2, \Delta_2^2\}$ with Δ_1 given in (16) and Δ_2 given in (19). We also obtain

$$\|\Delta B\|_\infty \leq C_1 \cdot \frac{\gamma \cdot (|\mathcal{S}| + K)^3}{\alpha^2 \cdot \xi \cdot \sigma \cdot \min\{\sigma^2, (1-\gamma)^2 \cdot \Delta\}} \cdot \frac{\log^2(N)}{N} \tag{133}$$

Therefore, as long as

$$C_1 \cdot \frac{(|\mathcal{S}| + K)^3}{\alpha^2 \cdot \xi \cdot \sigma \cdot \min\{\sigma^2, (1-\gamma)^2 \cdot \Delta\}} \cdot \frac{\log^2(N)}{N} \leq 1/\|(B^{\pi^*})^{-1}\|_\infty = 1/\sigma', \tag{134}$$

following standard perturbation analysis of linear equations [28], we have that

$$\frac{\|\boldsymbol{V}_r(\bar{\pi}^N) - \boldsymbol{V}_r(\pi^*)\|_\infty}{\|\boldsymbol{V}_r(\pi^*)\|_\infty} \leq C_2 \cdot \kappa(B^{\pi^*}) \cdot \left(\frac{\|\Delta B\|_\infty}{\|B^{\pi^*}\|_\infty} + \frac{\|\Delta \hat{\boldsymbol{r}}\|_\infty}{\|\hat{\boldsymbol{r}}(\pi^*)\|_\infty} \right), \tag{135}$$

where $\kappa(B^{\pi^*}) = \|B^{\pi^*}\|_\infty \cdot \|(B^{\pi^*})^{-1}\|_\infty$ denotes the conditional number of B^{π^*}, and C_2 is a constant. Note that we the regret

$$\begin{aligned} \text{Regret}_r(\bar{\pi}^N, N) &= \boldsymbol{\mu}^\top (\boldsymbol{V}_r(\bar{\pi}^N) - \boldsymbol{V}_r(\pi^*)) \leq (1-\gamma)\|\boldsymbol{V}_r(\bar{\pi}^N) - \boldsymbol{V}_r(\pi^*)\|_\infty \\ &\leq C_2(1-\gamma) \cdot \kappa(B^{\pi^*}) \cdot \|\boldsymbol{V}_r(\pi^*)\|_\infty \cdot \left(\frac{\|\Delta B\|_\infty}{\|B^{\pi^*}\|_\infty} + \frac{\|\Delta \hat{\boldsymbol{r}}\|_\infty}{\|\hat{\boldsymbol{r}}(\pi^*)\|_\infty} \right). \end{aligned} \tag{136}$$

It is clear to see that

$$\|\boldsymbol{V}_r(\pi^*)\|_\infty \leq \frac{1}{1-\gamma} \tag{137}$$

and

$$\|\boldsymbol{V}_r(\pi^*)\|_\infty \leq \frac{\|\hat{\boldsymbol{r}}(\pi^*)\|_\infty}{1-\gamma}. \tag{138}$$

Following [37], we have the following bound.

$$\sigma' = \|(B^{\pi^*})^{-1}\|_\infty \leq \frac{1}{1-\gamma}. \tag{139}$$

Also, from the definition, we have that

$$\|B^{\pi^*}\|_\infty = 1 - \gamma. \tag{140}$$

Plugging the bound (132), (133), (137), and (138), into the inequality (136), we have that

$$\text{Regret}_r(\bar{\pi}^N, N) \leq C_3 \cdot \frac{(|\mathcal{S}| + K)^3}{\alpha^2 \cdot \xi \cdot \sigma(1-\gamma) \cdot \min\{\sigma^2, (1-\gamma)^2 \cdot \Delta\}} \cdot \frac{\log^2(N)}{N} \tag{141}$$

where C_3 is a constant. In a same way, for any $k \in [K]$, we obtain that

$$\text{Regret}_k(\bar{\pi}^N, N) \leq C_3 \cdot \frac{(|\mathcal{S}| + K)^3}{\alpha^2 \cdot \xi \cdot \sigma(1-\gamma) \cdot \min\{\sigma^2, (1-\gamma)^2 \cdot \Delta\}} \cdot \frac{\log^2(N)}{N}. \tag{142}$$

To show the sample complexity bound, we denote by ε a constant such that

$$\varepsilon = C_3 \cdot \frac{(|\mathcal{S}| + K)^3}{\alpha^2 \cdot \xi \cdot \sigma(1-\gamma) \cdot \min\{\sigma^2, (1-\gamma)^2 \cdot \Delta\}} \cdot \frac{\log^2(N)}{N}.$$

Therefore, we have

$$N = O\left(\frac{(|\mathcal{S}| + K)^3}{\alpha^2 \cdot \xi \cdot \sigma(1-\gamma) \cdot \min\{\sigma^2, (1-\gamma)^2 \cdot \Delta\}} \cdot \frac{\log^2(1/\varepsilon)}{\varepsilon} \right).$$

Note that in each of the N rounds, we obtain a sample for each $(s, a) \in \mathcal{S} \times \mathcal{A}$. Therefore, the bound on N above should be multiplied by $|\mathcal{S}| \cdot |\mathcal{A}|$ to obtain the final sample complexity bound. Our proof is thus completed.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions and scope is accurately reflected.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the conclusion.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions and proofs are provided in full.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental details given in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: This is a theoretical paper and all the data is simulated with details given in the appendix.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the experimental details are given in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiment has been repeated enough times to guarantee that the expected regret has been appropriately approximated.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The simulated experiments are easy to implement and no significant computing resource is required.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: It conforms.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Presented in the conclusion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Not applied here.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Not applied here.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Not applied here.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Not applied here.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: None is included.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.