Learning Infinite-Horizon Average-Reward Markov Decision Processes with Constraints

Liyu Chen [1] **Rahul Jain** [1] **Haipeng Luo** [1]

Abstract

We study regret minimization for infinite-horizon average-reward Markov Decision Processes (MDPs) under cost constraints. We start by designing a policy optimization algorithm with carefully designed action-value estimator and bonus term, and show that for ergodic MDPs, our algorithm ensures $\tilde{\mathcal{O}}(\sqrt{T})$ regret and constant constraint violation, where T is the total number of time steps. This strictly improves over the algorithm of (Singh et al., 2020), whose regret and constraint violation are both $\tilde{\mathcal{O}}(T^{2/3})$. Next, we consider the most general class of weakly communicating MDPs. Through a finite-horizon approximation, we develop another algorithm with $\tilde{\mathcal{O}}(T^{2/3})$ regret and constraint violation, which can be further improved to $\tilde{\mathcal{O}}(\sqrt{T})$ via a simple modification, albeit making the algorithm computationally inefficient. As far as we know, these are the first set of provable algorithms for weakly communicating MDPs with cost constraints.

1. Introduction

Standard reinforcement learning (RL) algorithms aim at finding the optimal policy that maximizes the accumulated reward in a Markov Decision Process (MDP). In many real-world applications, however, the algorithm is also required to satisfy certain constraints. For example, in autonomous driving, the vehicle needs to reach the destination with minimum amount of time while obeying the traffic rules. These constrained versions of RL problems can be formulated by Constrained Markov Decision Processes (CMDPs) (Altman, 1999), where a learning agent tries to maximize the accumulated reward while ensuring that certain cost constraint is not violated or at least the violation is small enough.

Learning in a CMDP is a long-standing topic, and there is a surge of interest in it in light of all other theoretical advances in RL. Almost all recent works on CMDP, however, focus on the simpler finite-horizon setting (Kalagarla et al., 2021; Efroni et al., 2020; Qiu et al., 2020; Liu et al., 2021b) or the discounted setting (Liang et al., 2017; Tessler et al., 2018; Chen et al., 2021d; Liu et al., 2021a; Ding et al., 2020; 2021). In contrast, learning CMDP in the infinite-horizon average-reward setting, where the learner-environment interaction never ends or resets and the goal is to achieve optimal long-term average reward under constraints, appears to be much more challenging. For example, (Zheng & Ratliff, 2020) makes the restricted assumptions that the transition kernel is known and an initial policy that satisfies the constraints and induces an ergodic Markov chain is given, but still only achieves $\tilde{\mathcal{O}}(T^{3/4})$ regret after T steps (with no constraint violation). Another recent work (Singh et al., 2020) considers the special class of ergodic CMDPs, but only achieves $\tilde{\mathcal{O}}(T^{2/3})$ regret and $\tilde{\mathcal{O}}(T^{2/3})$ cost constraint violation. These existing results are far from optimal, exhibiting unique challenges of the constrained infinite-horizon average-reward setting.

In this work, we manage to overcome some of these challenges and significantly improve our understanding of regret minimization for infinite-horizon average-reward CMDPs. Our contributions are as follows:

- Following (Singh et al., 2020), we start by considering ergodic CMDPs in Section 3. We develop an algorithm that achieves $\tilde{\mathcal{O}}(\sqrt{T})$ regret and constant constraint violation, strictly improving (Singh et al., 2020). The main technical challenge in getting $\tilde{\mathcal{O}}(\sqrt{T})$ regret for the upper-confidence-type algorithm of (Singh et al., 2020) is the lack of a tighter bound on the span of the estimated bias function. Instead, we resolve this issue using a policy optimization algorithm with a special action-value estimator whose span is well controlled. To further control the transition estimation error from the action-value estimator, we also include a new bonus term in the policy update.
- In Section 4, we drop the ergodic assumption and consider the most general class of weakly communicating CMDPs. By reducing the original infinite-horizon problem to a finite-horizon problem similarly to (Wei et al., 2021), we show that a simple and efficient linear programming

[1]University of Southern California. Correspondence to: Liyu Chen <liyuc@usc.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

approach gives $\tilde{\mathcal{O}}(T^{2/3})$ regret and $\tilde{\mathcal{O}}(T^{2/3})$ constraint violation. Further introducing extra constraints to the linear program to control the span of some bias function, we are also able to obtain $\tilde{\mathcal{O}}(\sqrt{T})$ regret and $\tilde{\mathcal{O}}(\sqrt{T})$ violation, with the price that the resulting program can no longer be solved computationally efficiently. As far as we know, these are the first results for weakly communicating CMDPs (see some caveats below).

Related Work As mentioned, learning in CMDP is heavily studied recently in other settings (see references listed earlier), but for the infinite-horizon average-reward setting, other than the two recent works discussed above (Zheng & Ratliff, 2020; Singh et al., 2020), we are only aware of the works (Agarwal et al., 2021a;b) which study the ergodic case. Unfortunately, their results appear to be wrong due to a technical mistake which sidesteps an important challenge for this problem on controlling the span of some bias function; see Appendix A for more details.[1]

Regret minimization for the infinite-horizon average-reward setting without constraints dates back to (Bartlett & Tewari, 2009; Jaksch et al., 2010) and was shown to be possible only when the MDP is at least weakly communicating. Numerous improvements have been discovered in recent years; see e.g. (Ortner, 2018; Fruit et al., 2018; Talebi & Maillard, 2018; Abbasi-Yadkori et al., 2019; Zhang & Ji, 2019; Wei et al., 2020; 2021). From a technical perspective, designing provable algorithms for the infinite-horizon average-reward setting, especially for the general class of weakly communicating MDPs, has always been more challenging than other settings. For example, optimal model-free algorithms remain unknown for this setting (Wei et al., 2020), but have been developed for the finite-horizon setting (Jin et al., 2018) and the discounted setting (Dong et al., 2020).

Apart from MDPs, researchers also study constrained multi-armed bandit problems, such as conservative bandits (Wu et al., 2016; Kazerouni et al., 2017; Garcelon et al., 2020) and bandits with safety constraints modeled by a cost function (similar to our setting) (Amani et al., 2019; Pacchiano et al., 2021; Liu et al., 2021c).

2. Preliminaries

An infinite-horizon average-reward CMDP model is defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, c, \tau, P)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $r \in [0,1]^{\mathcal{S}\times\mathcal{A}}$ is the reward function, $c \in [0,1]^{\mathcal{S}\times\mathcal{A}}$ is the cost function modeling constraints, τ is a cost threshold, and $P = \{P_{s,a}\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ with $P_{s,a} \in \Delta_{\mathcal{S}}$ is the transition function, where $\Delta_{\mathcal{S}}$ is the simplex over $\mathcal{S}$. For simplicity, we assume that only the transition function P is unknown, while all other parameters are known. Dealing with unknown reward and cost functions can be done in a way similar to (Liu et al., 2021b) by maintaining standard confidence sets.

Throughout, we also assume that the MDP is *weakly communicating*, which is known to be necessary for learning even without constraints (Bartlett & Tewari, 2009). More specifically, an MDP is weakly communicating if its state space consists of two subsets: in the first subset, all states are transient under any stationary policy (that is, a mapping from $\mathcal{S}$ to $\Delta_{\mathcal{A}}$); in the second subset, every two state are communicating under some stationary policy.

The learning protocol is as follows: the learner starts from an arbitrary state $s_1 \in \mathcal{S}$, and interacts with the environment for T steps. In the t-th step, the learner observes state $s_t \in \mathcal{S}$, takes an action $a_t \in \mathcal{A}$, and transits to the next state $s_{t+1} \sim P_{s_t,a_t}$. Informally, the goal of the learner is to ensure large reward while at the same time incurring small cost relative to the threshold τ.

To describe these objectives formally, we introduce the concept of average utility function: for a stationary policy $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{S}}$, transition function P, and utility function $d \in \mathbb{R}_+^{\mathcal{S}\times\mathcal{A}}$, define the average utility for any $s \in \mathcal{S}$ as

$$J^{\pi,P,d}(s) = \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} d(s_t,a_t)\middle| \pi, P, s_1 = s\right]$$

where the expectation is with respect to the random sequence $a_1, s_2, a_2, s_3, a_3, \ldots$ generated according to $a_t \sim \pi$ and $s_{t+1} \sim P_{s_t,a_t}$. (Puterman, 1994, Theorem 8.3.2) shows that, there exists an optimal policy $\pi^\star$ such that for any $s \in \mathcal{S}$, $\pi^\star$ is the solution for the following optimization problem

$$\operatorname*{argmax}_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{S}}} J^{\pi,P,r}(s), \quad \text{s.t. } J^{\pi,P,c}(s) \le \tau, \tag{1}$$

and also $J^{\pi^\star,P,r}(s) = J^\star$ and $J^{\pi^\star,P,c}(s) = J_c^\star$ for some constants $J^\star$ and $J_c^\star$ independent of s. The performance of the learner is then measured through two quantities: first, her regret in reward against the optimal policy $\pi^\star$, defined as $R_T = \sum_{t=1}^{T}(J^\star - r(s_t,a_t))$, and second, her regret in cost against the threshold τ, or simply her constraint violation, defined as $C_T = \sum_{t=1}^{T}(c(s_t,a_t) - \tau)$.

Finally, (Puterman, 1994, Theorem 8.2.6) also shows that for any utility function d, there exists a *bias function* $q^{\pi,P,d} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ satisfying the Bellman equation: $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$,

$$q^{\pi,P,d}(s,a) + J^{\pi,P,d}(s) = d(s,a) + \mathbb{E}_{s'\sim P_{s,a}}[v^{\pi,P,d}(s')], \tag{2}$$

where $v^{\pi,P,d}(s) = \sum_{a\in\mathcal{A}} \pi(a|s) q^{\pi,P,d}(s,a)$, and also $J^{\pi,P,q^{\pi,P,d}}(s) = 0$ for all $s \in \mathcal{S}$. The functions q and

[1] However, the latest version of (Agarwal et al., 2021a) provides a fix to this issue and claims to achieve $\tilde{\mathcal{O}}(\frac{1}{\tau-c^0} LS\sqrt{AT})$ regret with constant constraint violation, where L is the maximum hitting time over all policies and all pair of source and destination states.

v are analogue of the well-known Q-function and state-value-function for the discounted or finite-horizon setting.

Notations Let $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$ be the number of states and actions respectively. For an integer n, $[n]$ denotes the set $\{1, \dots, n\}$. For a distribution $P \in \Delta_{\mathcal{S}}$ and a function $V \in \mathbb{R}^{\mathcal{S}}$, define $PV = \sum_{s \in \mathcal{S}} P(s)V(s)$. For any function $v \in \mathbb{R}^{\mathcal{S}}$, define its span as $\mathrm{sp}(v) = \max_{s \in \mathcal{S}} v(s) - \min_{s \in \mathcal{S}} v(s)$. When there is no confusion, we write $J^{\pi,P,d}$ as $J^{\pi,d}$, $q^{\pi,P,d}$ as $q^{\pi,d}$, and $v^{\pi,P,d}$ as $v^{\pi,d}$. Given a policy π and a transition P, define matrix P^π such that $P^\pi_{s,s'} = \sum_a \pi(a|s) P_{s,a}(s')$. For any $\epsilon \in (0,1)$, $\pi^{\star,\epsilon}$ is defined in the same way as $\pi^\star$ but with the threshold τ replaced by $\tau - \epsilon$. Let $J^{\star,\epsilon}$ denote the corresponding average reward $J^{\pi^{\star,\epsilon},r}(s)$ (that is s-independent as mentioned).

3. Results for Ergodic MDPs

We start by considering a special case of ergodic MDPs, which are self-explorative and often easier to learn compared to the general case of weakly communicating MDPs. However, even in this special case, the presence of cost constraint already makes the problem highly challenging as discussed below.

Specifically, an MDP is ergodic if for any stationary policy, the induced Markov chain is ergodic (that is, irreducible and aperiodic). There are several nice properties about ergodic MDPs. First, the long term average behavior of any stationary policy π is independent of the starting state: one can define the occupancy measure (also called stationary distribution) $\mu_{\pi,P} \in [0,1]^{\mathcal{S} \times \mathcal{A}}$ such that $\mu_{\pi,P}(s,a) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}[\sum_{t=1}^{T} \mathbb{I}\{s_t = s, a_t = a\} | \pi, P]$ is the fraction of visits to (s,a) in the long run following π in an ergodic MDP with transition P (the starting state is irrelevant to this value). We also write $\mu_{\pi,P}$ as μ_π when there is no confusion. By definition, for any utility function $d \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$, the average utility $J^{\pi,d}(s)$ is also s-independent and can be written as $\langle \mu_\pi, d \rangle$, denoted by $J^{\pi,d}$ for short.

Moreover, ergodic MDPs have finite *mixing time* and *hitting time*, defined as follows:

$$t_{\text{mix}} = \max_\pi \min \left\{ t \geq 1 : \left\| (P^\pi)^t_{s,\cdot} - \mu_\pi \right\|_1 \leq \frac{1}{4}, \forall s \right\},$$
$$t_{\text{hit}} = \max_\pi \max_s \frac{1}{\mu_\pi(s)},$$

where $\mu_\pi(s) = \sum_a \mu_\pi(s,a)$. In words, mixing time is the maximum time required for any policy starting at any initial state to make the state distribution $\frac{1}{4}$-close to its stationary distribution, and hitting time is the maximum inverse stationary probability of visiting any state under any policy. As in previous work, we assume that t_{mix}, t_{hit} are known, and T is large enough so that $T \geq 30A \max\{t_{\text{mix}}, t_{\text{hit}}\}$.

Compared to the finite-horizon setting, one key challenge for learning infinite-horizon average-reward MDPs with constraints is to control the span of the bias function with respect to some estimated transition. How to do so is highly unclear even though the same under the true transition is simply bounded by $\mathcal{O}(t_{\text{mix}})$ for an ergodic MDP. In fact, the MDP associated with the estimated transition might not be ergodic any more. In the seminal work of (Jaksch et al., 2010) for the unconstrained problem, they show that the span of the optimistic bias function is upper bounded by the diameter of the MDP. Their arguments, however, are not applicable when constraints are presented. This brings severe difficulties in the analysis for natural optimism-based approaches. For example, Singh et al. (2020) exploit the self-explorative property of ergodic MDPs to analyze an extension of the UCRL algorithm (Jaksch et al., 2010), but only manage to obtain $\tilde{\mathcal{O}}(T^{2/3})$ bounds for both regret R_T and constraint violation C_T. Moreover, their analysis does not generalize to the case of weakly communicating MDPs.

3.1. Our Algorithm

To resolve the issue mentioned above, we take a different approach — we adopt and extend the *policy optimization* algorithm of (Wei et al., 2020) (called MDP-OOMD) from the unconstrained setting to the constrained one. The advantage of policy optimization is that, instead of finding optimistic policies and transitions based on full planning, it updates policy incrementally based on an estimate of the current policy's bias function, which avoids the need to control the span of bias functions under estimated transition. This, however, requires a careful design of the estimate of the current policy's bias function, which is our key algorithmic novelty.

We start by describing the framework of our algorithm, which is similar to (Wei et al., 2020), and then highlight what the key differences are. The complete pseudocode is presented in Algorithm 1. Specifically, the algorithm proceeds in episodes of $H = \tilde{\mathcal{O}}(t_{\text{mix}} t_{\text{hit}})$ steps, with a total of $K = \frac{T}{H}$ episodes (assumed to be an integer for simplicity). In each episode k, the algorithm (Line 1) executes the same policy π_k for the entire H steps, collecting a trajectory $\mathcal{T}_k$ of the form $(s_{t_1}, a_{t_1}, \dots, s_{t_2}, a_{t_2})$ for $t_1 = (k-1)H + 1$ and $t_2 = kH$. Then, using this trajectory (together with other statistics), the algorithm (Line 2) invokes a procedure EstimateQ to compute a bias function estimator $\widehat{\beta}_k \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$, such that $\widehat{\beta}_k(s,a)$ approximately tells us how good taking action a at state s and then following π_k in the future is. With such an estimator, the algorithm (Line 3) updates the policy and find π_{k+1} for the next episode, using the classic Online Mirror Descent (OMD) framework. Below, we flesh out the details of each part.

Bias Function Estimates To simultaneously take reward and cost constraint into account, we adopt the common

Algorithm 1 Policy Optimization for Ergodic CMDP

Parameter: episode length H, number of episodes $K = T/H$, interval length N, learning rate θ, scaling parameter η, dual variable upper bound λ, cost slack ϵ; see Eq. (5).
Initialize: $\pi_1(a|s) = 1/A$ for $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\lambda_1 = 0$.
for $k = 1, \ldots, K$ **do**

1 Execute π_k for H steps and obtain trajectory $\mathcal{T}_k$.

2 $\widehat{\beta}_k = \textsc{EstimateQ}(\mathcal{T}_k, \bar{P}_k, r - \frac{\lambda_k}{\eta} c)$ (Algorithm 2) where $\bar{P}_k$ is empirical transition (3).

for *all* $s \in \mathcal{S}$ **do**

3 Update policy:

$$\pi_{k+1}(\cdot|s) = \operatorname*{argmax}_{\pi \in \bar{\Delta}} \left\{ \left\langle \pi, \widehat{\beta}_k + u_k \right\rangle - D(\pi, \pi_k) \right\}$$

where $D(p,q) = \frac{1}{\theta} \sum_a (p(a) \ln \frac{p(a)}{q(a)} - p(a) + q(a))$, $\bar{\Delta} = \Delta_{\mathcal{A}} \cap [\frac{1}{T}, 1]^{\mathcal{A}}$, and u_k is defined in Appendix B.6 (see also Section 3.1).

4 Update dual variable

$$\lambda_{k+1} = \min \left\{ \lambda, \max \left\{ 0, \lambda_k + \widehat{J}_k + \epsilon - \tau \right\} \right\},$$

where $\widehat{J}_k = \frac{1}{H-N} \sum_{h=N+1}^{H} c(s_h^k, a_h^k)$ with $s_h^k = s_{(k-1)H+h}$ and $a_h^k = a_{(k-1)H+h}$.

primal-dual approach (Efroni et al., 2020; Liu et al., 2021b) and consider the adjusted reward function $d = r - \frac{\lambda_k}{\eta} c$ for episode k, where η is a scaling parameter and λ_k is a dual variable (whose update will be discussed later). Intuitively, this adjusted reward provides a balance between maximizing rewards and minimizing costs. The procedure EstimateQ is effectively trying to estimate the bias function associated with the current policy π_k and the adjusted reward d, that is, $q^{\pi_k, d}$ (up to some additive term that is the same across all (s,a) entries), using data from the trajectory $\mathcal{T}_k$ of this episode. The pseudocode of EstimateQ in shown in Algorithm 2. It shares a similar framework as (Wei et al., 2020): for each $s \in \mathcal{S}$, it collects data from non-overlapping intervals of length $N = \tilde{\mathcal{O}}(t_{\text{mix}})$ that start from state s, and also make sure that these intervals are at least N steps apart from each other to reduce the correlation of data (see the while-loop of Algorithm 2).

However, different from (Wei et al., 2020) which uses standard importance weighting when constructing the estimator, we propose a new method that is critical to our analysis (see last two lines of Algorithm 2). Specifically, for each interval i mentioned above, we compute the cumulative adjusted reward y_i, and then average them over all intervals starting from state s as a value-function estimate $V(s)$. Finally, we return the bias function estimate whose (s,a) entry is $d(s,a) + \bar{P}_k V$, in light of the right-hand side of the Bellman

Algorithm 2 EstimateQ

Input: trajectory $\mathcal{T} = (s_{t_1}, a_{t_1}, \ldots, s_{t_2}, a_{t_2})$, empirical transition $\bar{P}$, and utility function d.
Define: $N = 4 t_{\text{mix}} \log_2 T$.
for *all* $s \in \mathcal{S}$ **do**
  Initialize: $\tau \leftarrow t_1, i \leftarrow 0$.
  while $\tau \le t_2 - N$ **do**
    if $s_\tau = s$ **then**
      $i \leftarrow i + 1$.
      $y_i = \sum_{t=\tau}^{\tau+N-1} d(s_t, a_t)$.
      $\tau \leftarrow \tau + 2N$.
    else $\tau \leftarrow \tau + 1$.
  Set $V(s) = \mathbb{I}\{i > 0\} \frac{1}{i} \sum_{j=1}^{i} y_j$.
return function Q such that $Q(s,a) = d(s,a) + \bar{P} V$.

equation (2). Here, $\bar{P}_k$ is the current empirical transition function such that

$$\bar{P}_{k,s,a}(s') = \frac{N_k(s,a,s')}{N_k^+(s,a)}, \tag{3}$$

where $N_k(s,a,s')$ is the number of visits to state-action-state triplet (s,a,s') before episode k and $N_k^+(s,a) = \max\{1, N_k(s,a)\}$ with $N_k(s,a) = \sum_{s'} N_k(s,a,s')$.

The reason of using this new estimator is to ensure that the final estimator $\widehat{\beta}_k(s,a)$ has a reasonable scale (roughly $\mathcal{O}(N) = \tilde{\mathcal{O}}(t_{\text{mix}})$), which in turn makes sure that the policy π_k is relatively stable across episodes. On the other hand, the importance-weighted estimator of (Wei et al., 2020) scales with $\frac{1}{\pi_k(a|s)}$ for the (s,a) entry, which could be very large. This is not an issue for their algorithm since they use a more stable regularizer called log-barrier in the OMD policy update, but this is not viable for us as explained next.

Policy Update With the estimate $\widehat{\beta}_k$, Algorithm 1 then updates π_k according to the OMD update: $\pi_{k+1}(\cdot|s) = \operatorname{argmax}_{\pi \in \bar{\Delta}} \left\{ \left\langle \pi, \widehat{\beta}_k + u_k \right\rangle - D(\pi, \pi_k) \right\}$. Here, $\bar{\Delta} = \Delta_{\mathcal{A}} \cap [\frac{1}{T}, 1]^{\mathcal{A}}$ is a truncated simplex, $D(p,q) = \frac{1}{\theta} \sum_a (p(a) \ln \frac{p(a)}{q(a)} - p(a) + q(a))$ is the KL-divergence scaled by the inverse of a learning rate $\theta > 0$, and finally u_k is an extra exploration bonus term used to cancel the bias due to using $\bar{P}_k$ instead of the true transition P when computing $\widehat{\beta}_k$. More concretely, u_k is an approximation of q^{π_k, P_k, x_k}, where x_k is a reward bonus function defined as

$$x_k(s,a) = \left(\frac{1}{\sqrt{N_k^+(s,a)}} + \sum_{a'} \frac{\pi_k(a'|s)}{\sqrt{N_k^+(s,a')}} \right) \iota, \tag{4}$$

for $\iota = \frac{2\lambda N}{\eta} \sqrt{S \ln \frac{2SAT}{\delta}}$ and failure probability $\delta \in (0,1)$, and P_k is an (approximate) optimistic transition with respect

to policy π_k and reward bonus x_k that lies in some transition confidence set. We compute P_k and u_k via Extended Value Iteration (EVI) (Jaksch et al., 2010) with precision $\epsilon_{\text{EVI}} = \frac{1}{T}$ as follows: we iteratively perform a value iteration procedure with $u^0(s) = 0$ and

$$u^{i+1}(s) = \sum_{a\in\mathcal{A}} \pi_k(a|s) \left(x_k(s,a) + \max_{P\in\mathcal{Q}_k} P_{s,a} u^i \right),$$

where $\mathcal{Q}_k$ is some transition confidence set formally defined in Appendix B.6. We stop the iteration above at index $i^\star$, which is the first index i such that $\text{sp}(u^{i+1} - u^i) \le \epsilon_{\text{EVI}}$. Then, we define $P_{k,s,a} = \text{argmax}_{P\in\mathcal{Q}_k} P_{s,a} u^{i^\star}$ as the approximate optimistic transition, and $u_k(s,a) = x_k(s,a) - \min_{s'} u^{i^\star+1}(s') + P_{k,s,a} u^{i^\star}$ as the shifted approximate bias function. The full setup is deferred to Appendix B.6.

If $\bar{\Delta}$ is replaced with the standard simplex $\Delta_{\mathcal{A}}$, then the update rule would be in the standard form of multiplicative weight update. However, we use a truncated simplex because our analysis requires a so-called "interval regret" guarantee from OMD (that is, regret measured on a specific interval; see the proof of Lemma 19), and multiplicative weight update over a truncated simplex is a standard way to achieve so. This is why we cannot use the log-barrier OMD of (Wei et al., 2020) because as far as we know it does not provide an interval regret guarantee.

Dual Variable Update The dual variable λ_{k+1} is updated via gradient descent $\lambda_k + \widehat{J}_k + \epsilon - \tau$, projected back to $[0, \lambda]$; see Line 4 of Algorithm 1. Here, ϵ is a cost slack, λ is an upper bound for the dual variable, and $\widehat{J}_k$ is the empirical average cost suffered by the learner in the last $H - N$ steps of this episode (discarding the first N steps to ensure that the state-action distribution is close to μ_{π_k} due to ergodicity). It can be shown that $\widehat{J}_k$ is an accurate estimate of $J^{\pi_k,c}$; see Lemma 7.

Parameter Tuning Finally, we list the exact value of the parameters below (recall $\iota = \frac{2\lambda N}{\eta}\sqrt{S \ln \frac{2SAT}{\delta}}$):

$$\begin{aligned}
&H = \lceil 16 t_{\text{mix}} t_{\text{hit}} (\log_2 T)^2 \rceil, \quad N = \lceil 4 t_{\text{mix}} \log_2 T \rceil, \\
&\theta = \min\left\{ 1/(4H\iota), \sqrt{\ln T/(4KH^2\iota^2)} \right\}, \\
&\eta = 1 + 2^{10}(\tau - c^0) N \sqrt{S} \ln\left(\frac{4SAT^3}{\delta}\right) \times \\
&\left(\sqrt{S^2AT} + \sqrt{HT} + S^{1.5}AH \ln\left(\frac{4SAT^3}{\delta}\right) \right), \\
&\lambda = \frac{40\eta}{\tau - c^0}, \quad \epsilon = \min\left\{ \frac{\tau - c^0}{2}, \frac{3\lambda}{K} \right\}.
\end{aligned} \tag{5}$$

Here, H and N are chosen such that Algorithm 2 gives accurate bias function estimates (Lemma 8); θ gives optimal regret bound trade-off for the PO analysis (Appendix B.4); λ and η are chosen such that the dual variables $\{\lambda_k\}_k$ fall in an appropriate range (Lemma 19), which is important to our analysis; ϵ is chosen such that the learner is just conservative enough to give constant constraint violation (Appendix B.5); c^0 is a constant such that there is a (strictly safe) policy π^0 with $c^0 = \max_s J^{\pi^0,P,c}(s) < \tau$. We assume that c^0 is known (but not the policy π^0), similarly to prior work such as (Efroni et al., 2020; Liu et al., 2021b).

3.2. Guarantees and Analysis

The guarantee of Algorithm 1 is summarized below.

Theorem 1. *With probability at least $1 - 16\delta$, Algorithm 1 ensures the following guarantees: $R_T = \tilde{\mathcal{O}}\left(\frac{t_{mix}^2 t_{hit}}{\tau - c^0} (\sqrt{S^3AT} + \sqrt{S t_{mix} t_{hit} T} + S^2 A t_{mix} t_{hit}) + \frac{t_{mix} t_{hit}}{(\tau - c^0)^2} \right)$ and $C_T = \tilde{\mathcal{O}}\left(\frac{t_{mix}^4 t_{hit}^2 S^3 A + t_{mix}^5 t_{hit}^3 S}{(\tau - c^0)^2} + \frac{t_{mix}^3 t_{hit}^2 S^2 A}{\tau - c^0} \right)$.*

Looking only at the dependence on T, our bounds are $R_T = \tilde{\mathcal{O}}(\sqrt{T})$ and $C_T = \tilde{\mathcal{O}}(1)$, improving the $\tilde{\mathcal{O}}(T^{2/3})$ bounds of (Singh et al., 2020) for both metrics. Below, we show a proof sketch of this theorem.

Regret For regret R_T, we start by decomposing it as

$$\begin{aligned}
R_T = \sum_{t=1}^{T} (J^\star - r(s_t, a_t)) = \underbrace{T(J^\star - J^{\star,\epsilon})}_{\text{DIFF}} + \\
\underbrace{H \sum_{k=1}^{K} (J^{\star,\epsilon} - J^{\pi_k, r})}_{\text{REG}} + \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} (J^{\pi_k, r} - r(s_h^k, a_h^k))}_{\text{DEV}},
\end{aligned}$$

where $s_h^k = s_{(k-1)H+h}$ and $a_h^k = a_{(k-1)H+h}$. We bound each of the three terms separately below.

Bounding DIFF DIFF is at most $\frac{\epsilon T}{\tau - c^0}$ by using the following lemma directly, which shows the difference in optimal reward when shrinking the constraint threshold by ϵ.

Lemma 1. *For $\epsilon \in [0, \tau - c^0]$, $J^\star - J^{\star,\epsilon} \le \frac{\epsilon}{\tau - c^0}$.*

Bounding REG We decompose the REG term as $\text{REG} = \sum_{k=1}^{K} (J^{\pi^{\star,\epsilon}, r - \frac{\lambda_k}{\eta} c} - J^{\pi_k, r - \frac{\lambda_k}{\eta} c}) + \sum_{k=1}^{K} \frac{\lambda_k}{\eta} (J^{\pi^{\star,\epsilon}, c} - J^{\pi_k, c})$. The first term can be rewritten by the value difference lemma (Lemma 16), and is bounded following the standard OMD analysis. The final result is shown below.

Lemma 2. *For any policy $\mathring{\pi}$ and a subset of episodes $\mathcal{I} = \{i, i+1, \ldots, j\} \subseteq [K]$, we have: $\sum_{k\in\mathcal{I}} J^{\mathring{\pi}, r - \frac{\lambda_k}{\eta} c} - J^{\pi_k, r - \frac{\lambda_k}{\eta} c} \le \frac{\lambda}{4(\tau - c^0)}$ with probability at least $1 - 4\delta$.*

For the second term, we have

$$\sum_{k=1}^{K}\frac{\lambda_k}{\eta}(J^{\pi^{\star,\epsilon},c}-J^{\pi_k,c}) \overset{\text{(i)}}{\leq} \sum_{k=1}^{K}\frac{\lambda_k}{\eta}\left(\tau-\epsilon-J^{\pi_k,c}\right)$$
$$\overset{\text{(ii)}}{\lesssim} \sum_{k=1}^{K}\frac{\lambda_k}{\eta}\left(\tau-\epsilon-\widehat{J}_k\right)+\sum_{k=1}^{K}\frac{\lambda_k}{\eta}(\widehat{J}_k-\mathbb{E}_k[\widehat{J}_k])$$
$$\overset{\text{(iii)}}{\leq} \frac{1}{\eta}\sum_{k=1}^{K}\lambda_k(\lambda_k-\lambda_{k+1})+\frac{\tau^2 K}{\eta}+\tilde{\mathcal{O}}\left(\frac{\lambda}{\eta}\sqrt{K}\right)$$
$$\overset{\text{(iv)}}{=} \tilde{\mathcal{O}}\left(\frac{K}{\eta}+\frac{\lambda}{\eta}\sqrt{K}\right) \overset{\text{(v)}}{=} \tilde{\mathcal{O}}\left(\frac{\lambda}{\tau-c^0}\right).$$

Here, (i) is by the definition of $\pi^{\star,\epsilon}$; (ii) is because $\mathbb{E}_k[\widehat{J}_k]$ is a good estimate of $J^{\pi_k,c}$ ($\mathbb{E}_k$ denotes the expectation given everything before episode k); (iii) applies Azuma's inequality and the following argument: if $\lambda_{k+1} > 0$, then $\tau-\epsilon-\widehat{J}_k \leq \lambda_k-\lambda_{k+1}$ by the definition of λ_k; otherwise, $\lambda_k \leq \tau-\epsilon-\widehat{J}_k < \tau$ and $\lambda_k(\tau-\epsilon-\widehat{J}_k) \leq \tau^2$; (iv) is because $\sum_{k=1}^{K}\lambda_k(\lambda_k-\lambda_{k+1}) = \frac{1}{2}\sum_{k=1}^{K}\left(\lambda_k^2-\lambda_{k+1}^2+(\lambda_{k+1}-\lambda_k)^2\right) \leq \frac{K}{2}$, where the last inequality is by $\lambda_1=0$ and $|\lambda_k-\lambda_{k+1}|\leq 1$; (v) is by the value of the parameters in Eq. (5). Putting everything together, we have $\textsc{Reg} = \tilde{\mathcal{O}}(\frac{\lambda}{\tau-c^0})$.

Bounding DEV We prove a more general statement saying that $\sum_{k=1}^{K}\sum_{h=1}^{H}(J^{\pi_k,d}-d(s_h^k,a_h^k)) \lesssim \lambda$ for any utility function $d \in [0,1]^{\mathcal{S}\times\mathcal{A}}$; see Lemma 18. The idea is as follows. Using the Bellman equation (2), the left-hand side is equal to $\sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s_h^k,a_h^k}v^{\pi_k,d}-q^{\pi_k,d}(s_h^k,a_h^k))$, which can then be decomposed as the sum of three terms: $\sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s_h^k,a_h^k}v^{\pi_k,d}-v^{\pi_k,d}(s_{h+1}^k))$, $\sum_{k=1}^{K}\sum_{h=1}^{H}(v^{\pi_k,d}(s_h^k)-q^{\pi_k,d}(s_h^k,a_h^k))$, and $\sum_{k=1}^{K}\sum_{h=1}^{H}(v^{\pi_k,d}(s_{h+1}^k)-v^{\pi_k,d}(s_h^k))$. The first two terms are sums of martingale difference sequence and of order $\tilde{\mathcal{O}}(t_{\text{mix}}\sqrt{T})$ by Azuma's inequality. The last term can be rearranged and telescoped to $\sum_{k=2}^{K}(v^{\pi_k,d}(s_1^k)-v^{\pi_{k-1},d}(s_1^k))$ (dropping two negligible terms). Now, this is exactly the term where we need the stability of π_k: as long as π_k changes slowly, this bound is sublinear in K. As discussed, we ensure this by using a new estimator $\widehat{\beta}_k$ whose scale is nicely bounded, allowing us to show the following.

Lemma 3. *For any k, we have $|\pi_k(a|s)-\pi_{k-1}(a|s)| \leq 8\theta H\iota\pi_{k-1}(a|s)$ and $\left|v^{\pi_k,d}(s)-v^{\pi_{k-1},d}(s)\right| \leq 65\theta HN^2\iota$ where $d\in[0,1]^{\mathcal{S}\times\mathcal{A}}$ is any utility function.*

Putting everything together then finishes the proof for R_T.

Constraint Violation For C_T, we decompose as: $C_T = \sum_{k=1}^{K}\sum_{h=1}^{H}(c(s_h^k,a_h^k)-J^{\pi_k,c})+H\sum_{k=1}^{K}(J^{\pi_k,c}-\tau)$. The first term is similar to DEV and is of order λ by Lemma 18. The second term is roughly $H\sum_{k=1}^{K}(\widehat{J}_k-\tau)$ (recall $\widehat{J}_k$ is a good estimator of $J^{\pi_k,c}$), and can be further bounded by $H(\lambda-K\epsilon)$. The reason of the last step is that in Lemma 19, using the interval regret property (Lemma 2) ensured by our OMD update, we show $\lambda_k<\lambda$, that is, the truncation at λ never happens in the update rule of λ_k (with high probability). This implies $\lambda_{k+1}\geq\lambda_k+\widehat{J}_k+\epsilon-\tau$ by definition, and rearranging thus shows $\sum_{k=1}^{K}(\widehat{J}_k-\tau) \leq \sum_{k=1}^{K}(\lambda_{k+1}-\lambda_k-\epsilon)=\lambda_{K+1}-K\epsilon\leq\lambda-K\epsilon$.

Now if $\epsilon=3\lambda/K$ (*c.f.* Eq. (5)), the negative term $-HK\epsilon$ above cancels out all the positive terms and $C_T\leq 0$. Otherwise, we have $T=\tilde{\mathcal{O}}(\frac{N^2H^2S^3A+N^2H^3S}{(\tau-c^0)^2}+\frac{NH^2S^2A}{\tau-c^0})$ by the definition of ϵ, and the trivial bound $C_T\leq T$ concludes the proof after we plug in the definition of N and H.

3.3. Experiments

We evaluate Algorithm 1 empirically on a variant of the single hop wireless network environment similar to (Singh et al., 2020), where a wireless node continuously transmits data packets to a receiver. The node consists of a queue of packets with finite capacity B. At time step t, the node needs to choose a transmission power $a_t\in\{0.1,0.9\}$ as action. Higher transmission power leads to higher probability of attempted transmission. The number of packets arriving at time step t is denoted by Y_t. We assume that $Y_t\in\{0,1,2,3\}$ are i.i.d. sampled, and the associated probability vector is $(0.65,0.2,0.1,0.05)$. The channel reliability is $p_r=0.9$, that is, each attempted transmission has a probability of 0.9 to succeed. Let Q_t be the length of the queue at time step t. Then, the dynamics of the queue length can be described as $Q_{t+1}=\min\{B,\max\{0,Q_t+Y_t-D_t\}\}$, where $D_t=1$ with probability $a_t\cdot p_r$ and $D_t=0$ otherwise. The goal of this network is to maintain a short queue length with minimum transmission power: at time step t, the reward is $r_t=1-a_t$ and the cost is $c_t=Q_t$. The per-round cost threshold τ is set to 3.

This setup is almost identical to the experiments of (Singh et al., 2020), except that their action a_t is binary, that is, chosen from $\{0,1\}$. However, as far as we know this makes the MDP non-ergodic (even though their results only hold for ergodic MDPs). We therefore changed it to $\{0.1,0.9\}$.

We run Algorithm 1 for $T=3\cdot10^6$ time steps and 5 different random seeds with the following manually best tuned parameters: $H=300$, $N=20$, $\eta=\sqrt{T}$, $\lambda=\eta$, $\epsilon=0.01$ and $\theta=10/\sqrt{T}$. We also scale ι and the range of transition confidence sets by a factor of $\sqrt{0.1}$ to accelerate learning. The cumulative regret and constraint violation are shown in Figure 1 with shaded area as 95% confidence interval. As predicted by our theory, Algorithm 1 achieves sub-linear regret growth while ensuring that the constraint violation is upper bounded by a constant. The oscillation of regret and constraint violation is due to the fact that dual

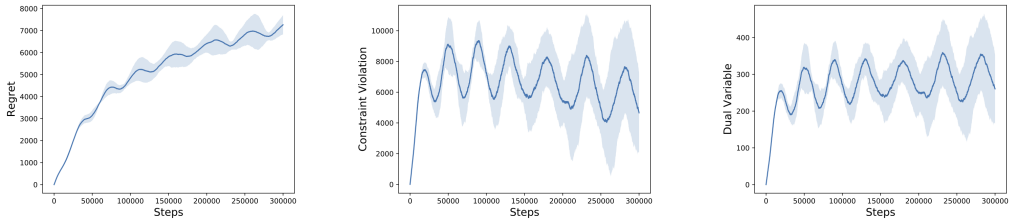

Figure 1. Experiment results of running Algorithm 1 on a variant of the single hop wireless network environment similar to (Singh et al., 2020). The plots from left to right are accumulated regret, accumulated constraint violation, and value of dual variables $\{\lambda_k\}_k$ in $3 \cdot 10^6$ time steps respectively. Each plot is an average of 5 repeated runs, and the shaded area is 95% confidence interval.

variables λ_k are changing adaptively, which controls how conservative the agent behaves with respect to the constraint violation. This is verified in the last plot of Figure 1 as well.

4. Results for Weakly Communicating MDPs

In this section, we drop the ergodic assumption and consider the most general case of weakly communicating MDPs. As in the unconstrained case, the span of the bias function of the optimal policy $\pi^\star$ plays an important role in this case and is unavoidable in the regret bound. More concretely, our bounds depend on $\mathrm{sp}(v^{\pi^\star,r})$ and $\mathrm{sp}(v^{\pi^\star,c})$, and our algorithm assumes knowledge of these two parameters, which we also write as $\mathrm{sp}_r^\star$ and $\mathrm{sp}_c^\star$ for short. We note that even in the unconstrained case, all existing algorithms whose regret bound depends on $\mathrm{sp}_r^\star$ need the knowledge of $\mathrm{sp}_r^\star$.

Weakly communicating MDPs impose extra challenges in learning. Specifically, there is no uniform bound for $\mathrm{sp}(v^{\pi,r})$ and $\mathrm{sp}(v^{\pi,c})$ for all stationary policy π (while in the ergodic case they are both $\tilde{\mathcal{O}}(t_{\mathrm{mix}})$). It is also unclear how to obtain an accurate estimate of a policy's bias function as in ergodic MDPs, which as we have shown is an important step for policy optimization algorithm. We are thus not able to extend the approach from Section 3 to this general case. Instead, we propose to solve the problem via a finite-horizon approximation, which is in spirit similar to another algorithm of (Wei et al., 2020) (called Optimistic Q-Learning).

Specifically, we still divide the T steps into K episodes, each of length H. In each episode, we treat it as an episodic finite-horizon MDP, and try to find a good (non-stationary) policy through the lens of occupancy measure, in which expected reward and cost are both linear functions and easy to optimize over. Concretely, consider a fixed starting state s, a non-stationary policy $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{S}\times[H]}$ whose behavior can vary in different steps within an episode, and an inhomogeneous transition function $P = \{P_{s,a,h}\}_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}$ where $P_{s,a,h} \in \Delta_{\mathcal{S}}$ specifies the probability of next state after taking action a at state s and step h. The corresponding occupancy measure $\nu_{\pi,P,s} \in [0,1]^{\mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{S}}$ is then such that $\nu_{\pi,P,s}(s',a,h,s'')$ is the probability of visiting state s' at step h, taking action a, and then transiting to state s'', if the learner starts from state s, executes π for the next H steps, and the transition dynamic follows P.

Conversely, a function $\nu \in [0,1]^{\mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{S}}$ is an occupancy measure with respect to a starting state s, some policy π_ν, and transition P_ν if and only if it satisfies:

1. Initial state is s: $\sum_a \sum_{s''} \nu(s',a,1,s'') = \mathbb{I}\{s'=s\}$.
2. Total mass is 1: $\sum_{s'}\sum_a\sum_{s''}\nu(s',a,h,s'') = 1,\ \forall h$.
3. Flow conservation: $\sum_{s''}\sum_a \nu(s'',a,h,s') = \sum_a\sum_{s''}\nu(s',a,h+1,s'')$ for all $s'\in\mathcal{S}, h\in[H-1]$.

We denote the set of all such ν as $\mathcal{V}_s$. For notational convenience, for a $\nu \in \mathcal{V}_s$, we define $\nu(s',a,h) = \sum_{s''}\nu(s',a,h,s'')$, $\nu(s',a) = \sum_h \nu(s',a,h)$, and $\nu(s',h) = \sum_a \nu(s',a,h)$. Also note that the corresponding policy π_ν and transition P_ν can be extracted via $\pi_\nu(a|s',h) = \frac{\nu(s',a,h)}{\nu(s',h)}$ and $P_{\nu,s',a,h}(s'') = \frac{\nu(s',a,h,s'')}{\nu(s',a,h)}$. These facts are taken directly from (Rosenberg & Mansour, 2019) (although there the starting state is always fixed). Note that $\mathcal{V}_s$ is a convex polytope with polynomial constraints. Also note that if one were to enforce P_ν to be homogeneous (that is, same across different steps of an episode), $\mathcal{V}_s$ would become non-convex. This is why we consider inhomogeneous transitions even though we know that the true transition is indeed homogeneous.

With the help of occupancy measure, we present two algorithms below. As far as we know, these are the first provable results for general weakly communicating CMDP.

4.1. An Efficient Algorithm

Our first algorithm is simple and computationally efficient; see Algorithm 3 for the pseudocode. At the beginning of each episode k, our algorithm observes the current state s_1^k and finds an occupancy measure $\nu_k \in \mathcal{V}_{k,s_1^k}$ that maximizes expected reward $\langle \nu, r\rangle = \sum_{s,a}\nu(s,a)r(s,a)$ under the cost constraint $\langle \nu, c\rangle \le H\tau + \mathrm{sp}_c^\star$. Here, $\mathcal{V}_{k,s_1^k}$ is a subset of

Algorithm 3 Finite Horizon Approximation for CMDP

Define: $H = \lceil (T/S^2A)^{1/3} \rceil$, $K = T/H$.
for $k = 1, \dots, K$ **do**
  Observe current state $s_1^k = s_{(k-1)H+1}$.
  Compute occupancy measure:

$$\nu_k = \operatorname*{argmax}_{\nu \in \mathcal{V}_{k,s_1^k} : \langle \nu, c \rangle \le H\tau + \mathrm{sp}_c^\star} \langle \nu, r \rangle , \tag{6}$$

  where $\mathcal{V}_{k,s} = \{\nu \in \mathcal{V}_s : P_\nu \in \mathcal{P}_k\}$ (see Eq. (7)).
  Extract policy $\pi_k = \pi_{\nu_k}$ from ν_k.
  for $h = 1, \dots, H$ **do**
    Play action $a_h^k \sim \pi_k(\cdot|s_h^k, h)$ and transit to s_{h+1}^k.

$\mathcal{V}_{s_1^k}$ such that P_ν for each $\nu \in \mathcal{V}_{k,s_1^k}$ lies in a standard Bernstein-type confidence set $\mathcal{P}_k$ defined as

$$\mathcal{P}_k = \Big\{ P' = \{P'_{s,a,h}\}_{(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]}, P'_{s,a,h} \in \Delta_{\mathcal{S}} : \\ \left|P'_{s,a,h}(s') - \bar{P}_{k,s,a}(s')\right| \\ \le 4\sqrt{\bar{P}_{k,s,a}(s')\alpha_k(s,a)} + 28\alpha_k(s,a) \Big\}, \tag{7}$$

where $\bar{P}_k$ is the same empirical transition function defined in Eq. (3), $\alpha_k(s,a) = \frac{\iota'}{N_k^+(s,a)}$, and $\iota' = \ln \frac{2SAT}{\delta}$. As a standard practice, $\mathcal{P}_k$ is constructed in a way such that it contains the true transition with high probability (Lemma 20). With ν_k, we simply follow the policy $\pi_k = \pi_{\nu_k}$ extracted from ν_k for the next H steps.

Note that the key optimization problem (6) in this algorithm can be efficiently solved, because the objective function is linear and the domain is again a convex polytope with polynomial constraints, thanks to the use of occupancy measures. We now state the main guarantee of Algorithm 3.

Theorem 2. *Algorithm 3 ensures with probability at least* $1 - 10\delta$,

$$R_T = \tilde{\mathcal{O}}\left((1 + \mathrm{sp}_r^\star)(S^2A)^{1/3}T^{2/3}\right), \\ C_T = \tilde{\mathcal{O}}\left((1 + \mathrm{sp}_c^\star)(S^2A)^{1/3}T^{2/3}\right).$$

Analysis As the first step, we need to quantify the finite-horizon approximation error. For any non-stationary policy $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{S}\times[H]}$, inhomogeneous transition function $P = \{P_{s,a,h}\}_{(s,a,h)}$, and utility function d, define value function $V_h^{\pi,P,d}(s) = \mathbb{E}\big[\sum_{h'=h}^H d(s_{h'}, a_{h'}) | \pi, P, s_h = s\big]$ where $a_{h'} \sim \pi(\cdot|s_{h'}, h')$ and $s_{h'+1} \sim P_{s_{h'},a_{h'},h'}$, and similarly action-value function $Q_h^{\pi,P,d}(s,a) = \mathbb{E}\big[\sum_{h'=h}^H d(s_{h'}, a_{h'}) | \pi, P, s_h = s, a_h = a\big]$. Additionally, define $V_{H+1}^{\pi,P,d}(s) = Q_{H+1}^{\pi,P,d}(s,a) = 0$. Further let $\widetilde{P} = \{\widetilde{P}_{s,a,h}\}_{(s,a,h)}$ be such that $\widetilde{P}_{s,a,h} = P_{s,a}$ (the true transition function) for all h. We often ignore the dependency on $\widetilde{P}$ and r for simplicity when there is no confusion. For example, V_h^π denotes $V_h^{\pi,\widetilde{P},r}$ and $V_h^{\pi,c}$ denotes $V_h^{\pi,\widetilde{P},c}$. For a stationary policy $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{S}}$, define $\widetilde{\pi}$ as the policy that mimics π in the finite-horizon setting, that is, $\widetilde{\pi}(\cdot|s,h) = \pi(\cdot|s)$. The following lemma shows the approximation error.

Lemma 4. *For any stationary policy* $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{S}}$ *and utility function* $d \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ *such that* $J^{\pi,d}(s) = J^{\pi,d}$ *for all* $s \in \mathcal{S}$, *we have* $|V_h^{\widetilde{\pi},d}(s) - (H - h + 1)J^{\pi,d}| \le \mathrm{sp}(v^{\pi,d})$ *for all state* $s \in \mathcal{S}$ *and* $h \in [H]$.

Next, in Lemma 22 (see appendix), we show that because $\mathcal{P}_k$ contains $\widetilde{P}$ with high probability, the occupancy measure with respect to the policy $\widetilde{\pi}^\star$ and the transition $\widetilde{P}$ is in the domain of (6) as well. As the last preliminary step, we bound the bias in value function caused by transition estimation, that is, difference between $\widetilde{P}$ and $P_k = P_{\nu_k}$.

Lemma 5. *For any utility function* $d \in [0,1]^{\mathcal{S}\times\mathcal{A}}$, *with probability at least* $1 - 4\delta$, *we have* $|\sum_{k=1}^K (V_1^{\pi_k,d}(s_1^k) - V_1^{\pi_k,P_k,d}(s_1^k))| = \tilde{\mathcal{O}}(\sqrt{S^2AH^2K} + H^2S^2A)$.

We are now ready to prove Theorem 2.

Proof of Theorem 2. We decompose R_T into three terms:

$$R_T = \sum_{t=1}^T J^\star - r(s_t, a_t) = \sum_{k=1}^K \left(HJ^\star - \sum_{h=1}^H r(s_h^k, a_h^k) \right) \\ = \sum_{k=1}^K \left(HJ^\star - V_1^{\widetilde{\pi}^\star}(s_1^k) \right) + \sum_{k=1}^K \left(V_1^{\widetilde{\pi}^\star}(s_1^k) - V_1^{\pi_k}(s_1^k) \right) \\ + \sum_{k=1}^K \left(V_1^{\pi_k}(s_1^k) - \sum_{h=1}^H r(s_h^k, a_h^k) \right).$$

The first term above is upper bounded by $K\mathrm{sp}_r^\star$ by Lemma 4. For the second term, by Lemma 22 and Lemma 5, we have

$$\sum_{k=1}^K \left(V_1^{\widetilde{\pi}^\star}(s_1^k) - V_1^{\pi_k}(s_1^k) \right) \le \sum_{k=1}^K \left(V_1^{\pi_k,P_k}(s_1^k) - V_1^{\pi_k}(s_1^k) \right) \\ = \tilde{\mathcal{O}}\left(\sqrt{S^2AH^2K} + H^2S^2A \right).$$

The last term is of order $\tilde{\mathcal{O}}(H\sqrt{K})$ by Azuma's inequality (Lemma 28). Using the definition of H and K, we arrive at

$$R_T = \tilde{\mathcal{O}}\left((1 + \mathrm{sp}_r^\star))(S^2A)^{1/3}T^{2/3}\right).$$

Algorithm 4 Finite Horizon Approximation for CMDP with Span Constraints

Define: $H = \sqrt{T/S^2A}$, $K = T/H$.
for $k = 1, \dots, K$ **do**
  Observe current state $s_1^k = s_{(k-1)H+1}$.
  Compute occupancy measure:

$$\nu_k = \operatorname*{argmax}_{\nu \in \mathcal{W}_{k,s_1^k}: \langle \nu, c \rangle \le H\tau + \mathrm{sp}_c^\star} \langle \nu, r \rangle, \tag{8}$$

  where $\mathcal{W}_{k,s} = \{\nu \in \mathcal{V}_s : P_\nu \in \mathcal{P}_k \text{ and } \forall h \in [H], \mathrm{sp}(V_h^{\nu,r}) \le 2\mathrm{sp}_r^\star, \mathrm{sp}(V_h^{\nu,c}) \le 2\mathrm{sp}_c^\star\}$.
  Extract policy $\pi_k = \pi_{\nu_k}$ from ν_k.
  for $h = 1, \dots, H$ **do**
    Play action $a_h^k \sim \pi_k(\cdot|s_h^k, h)$ and transit to s_{h+1}^k.

For constraint violations, we decompose C_T as:

$$\begin{aligned}\sum_{t=1}^T (c(s_t,a_t) - \tau) &= \sum_{k=1}^K \left(\sum_{h=1}^H c(s_h^k, a_h^k) - V_1^{\pi_k,c}(s_1^k)\right) \\ &+ \sum_{k=1}^K \left(V_1^{\pi_k,c}(s_1^k) - V_1^{\pi_k,P_k,c}(s_1^k)\right) \\ &+ \sum_{k=1}^K \left(V_1^{\pi_k,P_k,c}(s_1^k) - H\tau\right).\end{aligned}$$

The first term is again of order $\tilde{\mathcal{O}}(H\sqrt{K})$ by Azuma's inequality. The second term is of order $\tilde{\mathcal{O}}\left(\sqrt{S^2AH^2K} + H^2S^2A\right)$ by Lemma 5. The third term is upper bounded by $K\mathrm{sp}_c^\star$ because of the constraint $\langle \nu, c \rangle \le H\tau + \mathrm{sp}_c^\star$ in the optimization problem (6). Using the definition of H and K, we get $C_T = \tilde{\mathcal{O}}\left((1 + \mathrm{sp}_c^\star)(S^2A)^{1/3}T^{2/3}\right)$. This completes the proof. □

4.2. An Improved but Inefficient Algorithm

The bottleneck of the last algorithm/analysis is that the span of value functions are bounded by H, which is T-dependent and leads to sub-optimal dependency on T eventually. Below, we present an inefficient variant that achieves $\tilde{\mathcal{O}}(\sqrt{T})$ bounds for both regret and constraint violation.

The only new ingredient compared to Algorithm 3 is to enforce a proper upper bound on the span of value functions. Specifically, for any occupancy measure ν and utility function d, define $V_h^{\nu,d} = V_h^{\pi_\nu, P_\nu, d}$. We then enforce constraints $\mathrm{sp}(V_h^{\nu_k,r}) \le 2\mathrm{sp}_r^\star$ and $\mathrm{sp}(V_h^{\nu_k,c}) \le 2\mathrm{sp}_c^\star$; see the new domain $\mathcal{W}_{k,s}$ in the optimization problem (8) of Algorithm 4. This new domain is generally non-convex, making it unclear how to efficiently solve this optimization problem.

Nevertheless, we show the following improved guarantees.

Theorem 3. *Algorithm 4 ensures with probability at least* $1 - 6\delta$, $R_T = \tilde{\mathcal{O}}(\mathrm{sp}_r^\star S\sqrt{AT})$ *and* $C_T = \tilde{\mathcal{O}}(\mathrm{sp}_c^\star S\sqrt{AT})$.

Analysis Similarly to proving Theorem 2, we first show in Lemma 23 that the occupancy measure with respect to the policy $\widetilde{\pi}^\star$ and the transition $\widetilde{P}$ is in the domain of (8). We will also need the following key lemma.

Lemma 6. *For some utility function* $d \in [0,1]^{\mathcal{S}\times\mathcal{A}}$, *suppose* $\mathrm{sp}(V_h^{\pi_k,P_k,d}) \le B$ *for all episodes* $k \in [K]$ *and* $h \in [H]$. *Then, with probability at least* $1 - 2\delta$,

$$\begin{aligned}&\left|\sum_{k=1}^K \left(V_1^{\pi_k,P_k,d}(s_1^k) - \sum_{h=1}^H d(s_h^k, a_h^k)\right)\right| \\ &= \tilde{\mathcal{O}}\left((B+1)S\sqrt{AT} + BHS^2A\right).\end{aligned}$$

Proof of Theorem 3. We decompose the regret as follows:

$$\begin{aligned}R_T &= \sum_{t=1}^T J^\star - r(s_t,a_t) = \sum_{k=1}^K \left(HJ^\star - V_1^{\widetilde{\pi}^\star}(s_1^k)\right) \\ &+ \sum_{k=1}^K \left(V_1^{\widetilde{\pi}^\star}(s_1^k) - \sum_{h=1}^H r(s_h^k, a_h^k)\right) \\ &\le K\mathrm{sp}_r^\star + \sum_{k=1}^K \left(V_1^{\pi_k,P_k}(s_1^k) - \sum_{h=1}^H r(s_h^k, a_h^k)\right). \\ &\qquad \text{(Lemma 4 and Lemma 23)}\end{aligned}$$

By Lemma 6 with $B = 2\mathrm{sp}_r^\star$ and the definition of H and K, we have shown $R_T = \tilde{\mathcal{O}}\left((\mathrm{sp}_r^\star + 1)S\sqrt{AT}\right)$.

For constraint violation, by Lemma 6 with $B = 2\mathrm{sp}_c^\star$, and $V_1^{\pi_k,P_k,c}(s_1^k) - H\tau \le \mathrm{sp}_c^\star$ due to the constraint $\langle \nu, c \rangle \le H\tau + \mathrm{sp}_c^\star$ in (8), we have

$$\begin{aligned}C_T &= \sum_{t=1}^T (c(s_t,a_t) - \tau) = \sum_{k=1}^K \left(V_1^{\pi_k,P_k,c}(s_1^k) - H\tau\right) \\ &+ \sum_{k=1}^K \left(\sum_{h=1}^H c(s_h^k, a_h^k) - V_1^{\pi_k,P_k,c}(s_1^k)\right) \\ &= \tilde{\mathcal{O}}\left((\mathrm{sp}_c^\star + 1)S\sqrt{AT}\right).\end{aligned}$$

This completes the proof. □

We leave the question of how to achieve the same $\tilde{\mathcal{O}}(\sqrt{T})$ results with an efficient algorithm as a key future direction.

Acknowledgements

LC thanks Chen-Yu Wei for helpful discussions. HL is supported by NSF Award IIS-1943607 and a Google Faculty Research Award. RJ's research was supported by the NSF awards ECCS-1810447 and CCF-1817212, and by ONR award N00014-20-1-2258.

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pp. 3692–3702, 2019.

Agarwal, M., Bai, Q., and Aggarwal, V. Concave utility reinforcement learning with zero-constraint violations. *arXiv preprint arXiv:2109.05439*, 2021a.

Agarwal, M., Bai, Q., and Aggarwal, V. Markov decision processes with long-term average constraints. *arXiv preprint arXiv:2106.06680*, 2021b.

Altman, E. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Amani, S., Alizadeh, M., and Thrampoulidis, C. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, 2019.

Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 35–42. AUAI Press, 2009.

Chen, L. and Luo, H. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. In *International Conference on Machine Learning*, 2021.

Chen, L., Jafarnia-Jahromi, M., Jain, R., and Luo, H. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *Advances in Neural Information Processing Systems*, 2021a.

Chen, L., Jain, R., and Luo, H. Improved no-regret algorithms for stochastic shortest path with linear mdp. *arXiv preprint arXiv:2112.09859*, 2021b.

Chen, L., Luo, H., and Wei, C.-Y. Impossible tuning made possible: A new expert algorithm and its applications. *International Conference on Algorithmic Learning Theory*, 2021c.

Chen, Y., Dong, J., and Wang, Z. A primal-dual approach to constrained markov decision processes. *arXiv preprint arXiv:2101.10895*, 2021d.

Cohen, A., Kaplan, H., Mansour, Y., and Rosenberg, A. Near-optimal regret bounds for stochastic shortest path. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 8210–8219. PMLR, 2020.

Cohen, A., Efroni, Y., Mansour, Y., and Rosenberg, A. Minimax regret for stochastic shortest path. *Advances in Neural Information Processing Systems*, 2021.

Ding, D., Zhang, K., Basar, T., and Jovanovic, M. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanovic, M. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3304–3312. PMLR, 2021.

Dong, K., Wang, Y., Chen, X., and Wang, L. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *International Conference on Learning Representations*, 2020.

Efroni, Y., Mannor, S., and Pirotta, M. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pp. 1573–1581, 2018.

Garcelon, E., Ghavamzadeh, M., Lazaric, A., and Pirotta, M. Improved algorithms for conservative exploration in bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3962–3969, 2020.

Hazan, E. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

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

Kalagarla, K. C., Jain, R., and Nuzzo, P. A sample-efficient algorithm for episodic finite-horizon MDP with constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Kazerouni, A., Ghavamzadeh, M., Abbasi-Yadkori, Y., and Van Roy, B. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, 2017.

Liang, Q., Que, F., and Modiano, E. Accelerated primal-dual policy optimization for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., and Tian, C. Fast global convergence of policy optimization for constrained mdps. *arXiv preprint arXiv:2111.00552*, 2021a.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., and Tian, C. Learning policies with zero or bounded constraint violation for constrained MDPs. In *Advances in Neural Information Processing Systems*, 2021b.

Liu, X., Li, B., Shi, P., and Ying, L. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. In *Advances in Neural Information Processing Systems*, 2021c.

Ortner, R. Regret bounds for reinforcement learning via Markov chain concentration. *arXiv preprint arXiv:1808.01813*, 2018.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 2827–2835. PMLR, 2021.

Puterman, M. L. Markov decision processes: Discrete stochastic dynamic programming, 1994.

Qiu, S., Wei, X., Yang, Z., Ye, J., and Wang, Z. Upper confidence primal-dual reinforcement learning for CMDP with adversarial loss. In *Advances in Neural Information Processing Systems*, 2020.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5478–5486, 2019.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 8604–8613, 2020.

Singh, R., Gupta, A., and Shroff, N. B. Learning in Markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

Talebi, M. S. and Maillard, O.-A. Variance-aware regret bounds for undiscounted reinforcement learning in MDPs. In *Algorithmic Learning Theory*, pp. 770–805, 2018.

Tessler, C., Mankowitz, D. J., and Mannor, S. Reward constrained policy optimization. In *International Conference on Learning Representations (ICLR)*, 2018.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *International Conference on Machine Learning*, pp. 10170–10180. PMLR, 2020.

Wei, C.-Y., Jahromi, M. J., Luo, H., and Jain, R. Learning infinite-horizon average-reward MDPs with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3007–3015. PMLR, 2021.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Wu, Y., Shariff, R., Lattimore, T., and Szepesvári, C. Conservative bandits. In *International Conference on Machine Learning*, pp. 1254–1262. PMLR, 2016.

Zhang, Z. and Ji, X. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, 2019.

Zheng, L. and Ratliff, L. Constrained upper confidence reinforcement learning. In *Learning for Dynamics and Control*, pp. 620–629. PMLR, 2020.

A. Preliminaries for the Appendix

Notations Note that all algorithms proposed in this paper divide T steps into K episodes. Throughout the appendix, denote by $\mathbb{E}_k[\cdot]$ the expectation conditioned on the events before episode k and define $P_h^k = P_{s_h^k, a_h^k}$.

Issues of Some Related Work In (Agarwal et al., 2021a), they bound the span of the bias function w.r.t learner's policy and some estimated transition function by diameter D; see their Equation (90). They directly quote (Jaksch et al., 2010) as the reasonings. However, the arguments in (Jaksch et al., 2010) is only applicable when the learner's policy is computed by Extended Value Iteration without constraints, while their learner's policy is computed by solving some constrained optimization problem.

In (Agarwal et al., 2021b, Lemma 11), they claim that the span of the bias function of any stationary policy is upper bounded by D, which is clearly not true. Again, their learner's policy is computed by solving some constrained optimization problem.

B. Omitted Details for Section 3

Notations Define function $\widehat{V}_k$ such that $\widehat{\beta}_k(s,a) = r(s,a) - \frac{\lambda_k}{\eta} c(s,a) + \bar{P}_{k,s,a}\widehat{V}_k$. Note that $\left\|\widehat{V}_k\right\|_\infty \le \frac{2\lambda N}{\eta}$. For any subset of episodes $\mathcal{I} = \{i, \dots, i+1, j\} \subseteq [K]$, define $\mathcal{I}_{[1]} = i$ as the smallest element in $\mathcal{I}$.

B.1. Proof of Lemma 1

Proof. Since the occupancy measure space is convex, one can find a policy π' such that $\mu_{\pi'} = (1-\gamma)\mu_{\pi^\star} + \gamma\mu_{\pi^0}$ with $\gamma = \frac{\epsilon}{\tau - c^0}$. Now by $J^{\pi,d} = \langle \mu_\pi, d\rangle$:

$$J^{\pi',c} = (1-\gamma)J^{\pi^\star,c} + \gamma J^{\pi^0,c} \le (1-\gamma)\tau + \gamma c^0 \le \tau - \epsilon.$$

Thus, $J^\star - J^{\star,\epsilon} \le J^\star - J^{\pi',r} = \gamma(J^\star - J^{\pi^0,r}) \le \frac{\epsilon}{\tau - c^0}$. □

B.2. Bounding Estimation Error of $\widehat{\beta}_k$ and $\widehat{J}_k$

The next two lemmas bound the bias of $\widehat{\beta}_k$ and $\widehat{J}_k$ w.r.t the quantities they estimate.

Lemma 7. $\left|\mathbb{E}_k[\widehat{J}_k] - J^{\pi_k,c}\right| \le 1/T^2$.

Proof. For $h \ge N$, we have $\left\|(P^{\pi_k})^h_{s,\cdot} - \mu_{\pi_k}\right\|_1 \le \frac{2}{T^4}$ for any $s \in \mathcal{S}$ by Lemma 11. Thus,

$$\left|\mathbb{E}_k[\widehat{J}_k] - J^{\pi_k,c}\right| = \left|\frac{1}{H-N}\sum_{h=N}^{H-1}\sum_{s'}((P^{\pi_k})^h_{s_1^k,s'} - \mu_{\pi_k}(s'))\sum_a \pi_k(a|s')c(s',a)\right| \le 1/T^2.$$

□

Lemma 8. $\left|\mathbb{E}_k[\widehat{V}_k(s)] - v^{\pi_k, r-\frac{\lambda_k}{\eta}c}(s) - NJ^{\pi_k, r-\frac{\lambda_k}{\eta}c}\right| \le \frac{\lambda}{\eta T}$.

Proof. Note that $\widehat{V}_k(s) = \sum_a \pi_k(a|s)\widehat{Q}_k(s,a)$ where $\widehat{Q}_k$ is the estimator $\widehat{\beta}_k$ in (Wei et al., 2020, Lemma 6) with reward $r - \frac{\lambda_k}{\eta}c$, and we have $|\mathbb{E}_k[\widehat{Q}_k(s,a)] - q^{\pi_k, r-\frac{\lambda_k}{\eta}c}(s,a) - NJ^{\pi_k, r-\frac{\lambda_k}{\eta}c}| \le \frac{30\lambda}{\eta T^2}$ (the constant is determined by tracing their proof of Lemma 6 in Appendix B.2) by $\left\|r - \frac{\lambda_k}{\eta}c\right\|_\infty \le 2\lambda/\eta$. Then by $T \ge 30\max\{t_{\text{mix}}, t_{\text{hit}}\}$ and $\sum_a \pi_k(a|s)q^{\pi_k, r-\frac{\lambda_k}{\eta}c}(s,a) = v^{\pi_k, r-\frac{\lambda_k}{\eta}c}(s)$, the statement is proved. □

B.3. Proof of Lemma 3

Proof. By the update rule of π_k and following the proof of (Chen et al., 2021c, Lemma 17) (by Lemma 15, we have $c_{\max} = 2H\iota$ in their proof), we have

$$\pi_k(a|s) \in [\exp(-4\theta H\iota), \exp(4\theta H\iota)]\,\pi_{k-1}(a|s).$$

Therefore, by $|e^x - 1| \le 2|x|$ for $x \in [-1, 1]$, we have $|\pi_k(a|s) - \pi_{k-1}(a|s)| \le 8\theta H\iota\pi_{k-1}(a|s)$. For the second statement, first note that by Lemma 16 and Lemma 13:

$$\begin{aligned}\left|J^{\pi_k,d} - J^{\pi_{k-1},d}\right| &= \left|\sum_{s,a}\mu_{\pi_k}(s)(\pi_k(a|s) - \pi_{k-1}(a|s))q^{\pi_{k-1},d}(s,a)\right| \le \sum_{s,a}\mu_{\pi_k}(s)\left|\pi_k(a|s) - \pi_{k-1}(a|s)\right|\left|q^{\pi_{k-1},d}(s,a)\right| \\ &\le \sum_{s,a}\mu_{\pi_k}(s)\cdot 8\theta H\iota\pi_{k-1}(a|s)\cdot 6t_{\text{mix}} \le 48\theta H\iota t_{\text{mix}}.\end{aligned}$$

Next, for any policy π, define $d^\pi(s) = \sum_a \pi(a|s)d(s,a)$. Note that by $\langle \mu_\pi, d^\pi\rangle = J^{\pi,d}$,

$$v^{\pi,d}(s) = \sum_{t=0}^{\infty}\left\langle (P^\pi)^t_{s,\cdot} - \mu_\pi, d^\pi\right\rangle = \sum_{t=0}^{N-1}\left\langle (P^\pi)^t_{s,\cdot}, d^\pi\right\rangle - NJ^{\pi,d} + \sum_{t=N}^{\infty}\left\langle (P^\pi)^t_{s,\cdot} - \mu_\pi, d^\pi\right\rangle.$$

Moreover, by Lemma 12, we have $\left|\sum_{t=N}^{\infty}\left\langle (P^\pi)^t_{s,\cdot} - \mu_\pi, d^\pi\right\rangle\right| \le \frac{1}{T^3}$ for any policy π. Therefore,

$$\begin{aligned}&|v^{\pi_k,d}(s) - v^{\pi_{k-1},d}(s)| \\ &\le \left|\sum_{t=0}^{N-1}\left\langle (P^{\pi_k})^t_{s,\cdot} - (P^{\pi_{k-1}})^t_{s,\cdot}, d^{\pi_k}\right\rangle\right| + \left|\sum_{t=0}^{N-1}\left\langle (P^{\pi_{k-1}})^t_{s,\cdot}, d^{\pi_k} - d^{\pi_{k-1}}\right\rangle\right| + N\left|J^{\pi_k,d} - J^{\pi_{k-1},d}\right| + \frac{2}{T^3} \\ &\le \sum_{t=0}^{N-1}\left\|(P^{\pi_k})^t - (P^{\pi_{k-1}})^t)d^{\pi_k}\right\|_\infty + \sum_{t=0}^{N-1}\left\|d^{\pi_k} - d^{\pi_{k-1}}\right\|_\infty + 48\theta HN\iota t_{\text{mix}} + \frac{2}{T^3}.\end{aligned}$$

For the first term, note that:

$$\begin{aligned}\left\|((P^{\pi_k})^t - (P^{\pi_{k-1}})^t)d^{\pi_k}\right\|_\infty &\le \left\|P^{\pi_k}((P^{\pi_k})^{t-1} - (P^{\pi_{k-1}})^{t-1})d^{\pi_k}\right\|_\infty + \left\|(P^{\pi_k} - P^{\pi_{k-1}})(P^{\pi_{k-1}})^{t-1}d^{\pi_k}\right\|_\infty \\ &\le \left\|((P^{\pi_k})^{t-1} - (P^{\pi_{k-1}})^{t-1})d^{\pi_k}\right\|_\infty + \max_s\left\|P^{\pi_k}_{s,\cdot} - P^{\pi_{k-1}}_{s,\cdot}\right\|_1. \\ &\qquad\qquad \text{(every row of } P^{\pi_k} \text{ sums to 1 and } \left\|(P^{\pi_{k-1}})^{t-1}d^{\pi_k}\right\|_\infty \le 1)\end{aligned}$$

Moreover, by $|\pi_k(a|s) - \pi_{k-1}(a|s)| \le 8\theta H\iota\pi_{k-1}(a|s)$,

$$\max_s\left\|P^{\pi_k}_{s,\cdot} - P^{\pi_{k-1}}_{s,\cdot}\right\|_1 = \max_s\left|\sum_{s'}\sum_a(\pi_k(a|s) - \pi_{k-1}(a|s))P_{s,a}(s')\right| \le 8\theta H\iota.$$

Plugging this back and by a recursive argument, we get $\left\|((P^{\pi_k})^t - (P^{\pi_{k-1}})^t)d^{\pi_k}\right\|_\infty \le 8\theta H\iota t$. Moreover, $\sum_{t=0}^{N-1}\left\|d^{\pi_k} - d^{\pi_{k-1}}\right\|_\infty \le \sum_{t=0}^{N-1}\max_s\left\|\pi_k(\cdot|s) - \pi_{k-1}(\cdot|s)\right\|_1 \le 8\theta HN\iota$. Thus,

$$\begin{aligned}|v^{\pi_k,d}(s) - v^{\pi_{k-1},d}(s)| &\le 8\theta HN^2\iota + 8\theta HN\iota + 48\theta HN\iota t_{\text{mix}} + 2/T^3 \le 65\theta HN^2\iota. \\ &\qquad (\left\|d^{\pi_k} - d^{\pi_{k-1}}\right\|_\infty \le \max_s|\textstyle\sum_a(\pi_k(a|s) - \pi_{k-1}(a|s))d(s,a)| \le 8\theta H\iota)\end{aligned}$$

This completes the proof of the second statement. □

B.4. Proof of Lemma 2

Proof. Define policy π such that $\pi(a|s) = (1 - \frac{A}{T})\mathring{\pi}(a|s) + \frac{1}{T}$. Clearly, $\pi \in \bar{\Delta}$. Moreover, by Lemma 16 and Lemma 13:

$$\begin{aligned}\sum_{k\in\mathcal{I}}(J^{\mathring{\pi},r-\frac{\lambda_k}{\eta}c} - J^{\pi,r-\frac{\lambda_k}{\eta}c}) &= \sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\mathring{\pi}(a|s) - \pi(a|s))q^{\mathring{\pi},r-\frac{\lambda_k}{\eta}c}(s,a) \\ &\le \sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)\left(\frac{A}{T}\mathring{\pi}(a|s) - \frac{1}{T}\right)q^{\mathring{\pi},r-\frac{\lambda_k}{\eta}c}(s,a) \le \frac{12A\lambda}{\eta}.\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{k\in\mathcal{I}}(J^{\mathring{\pi},r-\frac{\lambda_k}{\eta}c}-J^{\pi_k,r-\frac{\lambda_k}{\eta}c})=\sum_{k\in\mathcal{I}}(J^{\mathring{\pi},r-\frac{\lambda_k}{\eta}c}-J^{\pi,r-\frac{\lambda_k}{\eta}c})+\sum_{k\in\mathcal{I}}(J^{\pi,r-\frac{\lambda_k}{\eta}c}-J^{\pi_k,r-\frac{\lambda_k}{\eta}c})\\
&\leq\frac{12A\lambda}{\eta}+\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))(q^{\pi_k,r-\frac{\lambda_k}{\eta}c}(s,a)+(N+1)J^{\pi_k,r-\frac{\lambda_k}{\eta}c})\\
&\qquad\qquad\text{(Lemma 16 and }\textstyle\sum_a(\pi(a|s)-\pi_k(a|s))(N+1)J^{\pi_k,r-\frac{\lambda_k}{\eta}c}=0)\\
&=\frac{12A\lambda}{\eta}+\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))\widehat{\beta}_k(s,a)\\
&\quad+\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))\left(P_{s,a}(v^{\pi_k,r-\frac{\lambda_k}{\eta}c}+NJ^{\pi_k,r-\frac{\lambda_k}{\eta}c})-\bar{P}_{k,s,a}\widehat{V}_k\right)\\
&=\frac{12A\lambda}{\eta}+\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))\widehat{\beta}_k(s,a)\\
&\quad+\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))P_{s,a}\left(v^{\pi_k,r-\frac{\lambda_k}{\eta}c}+NJ^{\pi_k,r-\frac{\lambda_k}{\eta}c}-\widehat{V}_k\right)\\
&\quad+\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))(P_{s,a}-\bar{P}_{k,s,a})\widehat{V}_k.
\end{aligned}$$

For the third term above, by Lemma 8, Lemma 28, and $\left\|\widehat{V}_k\right\|_\infty\leq\frac{2\lambda N}{\eta}$, with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))P_{s,a}\left(v^{\pi_k,r-\frac{\lambda_k}{\eta}c}+NJ^{\pi_k,r-\frac{\lambda_k}{\eta}c}-\widehat{V}_k\right)\\
&\leq\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))P_{s,a}\left(\mathbb{E}_k[\widehat{V}_k]-\widehat{V}_k\right)+\frac{\lambda}{\eta}\leq\frac{4\lambda N}{\eta}\sqrt{|\mathcal{I}|\ln\frac{4T^3}{\delta}}.
\end{aligned}$$

For the fourth term above, with probability at least $1-\delta$:

$$\begin{aligned}
&\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))(P_{s,a}-\bar{P}_{k,s,a})\widehat{V}_k\leq\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\pi(a|s)+\pi_k(a|s))\left\|P_{s,a}-\bar{P}_{k,s,a}\right\|_1\left\|\widehat{V}_k\right\|_\infty\\
&\qquad\qquad\text{(Cauchy-Schwarz inequality)}\\
&\leq\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s,a)\frac{2\lambda N}{\eta}\sqrt{S\ln\frac{2SAT}{\delta}}\left(\frac{1}{\sqrt{N_k^+(s,a)}}+\sum_{a'}\frac{\pi_k(a'|s)}{\sqrt{N_k^+(s,a')}}\right)\\
&\qquad\qquad\text{(Lemma 30 with a union bound over }\mathcal{S}\times\mathcal{A}\times[T]\text{ and }\left\|\widehat{V}_k\right\|_\infty\leq\tfrac{2\lambda N}{\eta})\\
&=\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s,a)x_k(s,a)=\sum_{k\in\mathcal{I}}J^{\pi,P,x_k}=\sum_{k\in\mathcal{I}}(J^{\pi,P,x_k}-J^{\pi_k,P_k,x_k})+\sum_{k\in\mathcal{I}}J^{\pi_k,P_k,x_k}.
\end{aligned}$$

By Lemma 10,

$$\begin{aligned}
J^{\pi,P,x_k}-J^{\pi_k,P_k,x_k}&\leq\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))u_k(s,a)+\sum_{s,a}\mu_\pi(s,a)(P_{s,a}-P_{k,s,a})u_k'+\epsilon_{\text{EVI}}\\
&\leq\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))u_k(s,a)+\epsilon_{\text{EVI}}.\qquad\text{(definition of }P_{k,s,a})
\end{aligned}$$

Substituting these back, we have:

$$\begin{aligned}
&\sum_{k\in\mathcal{I}}J^{\mathring{\pi},r-\frac{\lambda_k}{\eta}c}-J^{\pi_k,r-\frac{\lambda_k}{\eta}c}\\
&\leq\frac{12A\lambda}{\eta}+\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))(\widehat{\beta}_k(s,a)+u_k(s,a))+\frac{4\lambda N}{\eta}\sqrt{|\mathcal{I}|\ln\frac{4T^3}{\delta}}+\sum_{k\in\mathcal{I}}J^{\pi_k,P_k,x_k}+K\epsilon_{\text{EVI}}.
\end{aligned}$$

Note that by the standard OMD analysis (Hazan, 2019), for any $s \in \mathcal{S}$:

$$\begin{aligned}&\sum_{k\in\mathcal{I}}\sum_{a\in\mathcal{A}}(\pi(a|s)-\pi_k(a|s))(\widehat{\beta}_k(s,a)+u_k(s,a))\\&\le\sum_{k\in\mathcal{I}}\left(D(\pi(\cdot|s),\pi_k(\cdot|s))-D(\pi(\cdot|s),\pi_{k+1}(\cdot|s))+D(\pi_k(\cdot|s),\pi'_{k+1}(\cdot|s))\right),\end{aligned}$$

where $\pi'_{k+1}(a|s)=\pi_k(a|s)\exp(\theta(\widehat{\beta}_k(s,a)+u_k(s,a)))$. Then by $\theta\left|\widehat{\beta}_k(s,a)+u_k(s,a)\right|\le 2\theta H\iota\le 1$ (Lemma 15):

$$\begin{aligned}D(\pi_k(\cdot|s),\pi'_{k+1}(\cdot|s))&=\frac{1}{\theta}\sum_{a\in\mathcal{A}}\left(\pi_k(a|s)\ln\frac{\pi_k(a|s)}{\pi'_{k+1}(a|s)}-\pi_k(a|s)+\pi'_{k+1}(a|s)\right)\\&=\frac{1}{\theta}\sum_{a\in\mathcal{A}}\pi_k(a|s)\left(-\theta(\widehat{\beta}_k(s,a)+u_k(s,a))-1+e^{\theta(\widehat{\beta}_k(s,a)+u_k(s,a))}\right)\\&\le\theta\sum_{a\in\mathcal{A}}\pi_k(a|s)\left(\widehat{\beta}_k(s,a)+u_k(s,a)\right)^2.&&(e^{-x}-1+x\le x^2\text{ for }x\ge-1)\end{aligned}$$

Therefore,

$$\begin{aligned}&\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_\pi(s)(\pi(a|s)-\pi_k(a|s))(\widehat{\beta}_k(s,a)+u_k(s,a))\\&\le\sum_{s\in\mathcal{S}}\mu_\pi(s)\sum_{k\in\mathcal{I}}\left(D(\pi(\cdot|s),\pi_k(\cdot|s))-D(\pi(\cdot|s),\pi_{k+1}(\cdot|s))+\theta\sum_{a\in\mathcal{A}}\pi_k(a|s)\left(\widehat{\beta}_k(s,a)+u_k(s,a)\right)^2\right)\\&\le\sum_{s\in\mathcal{S}}\mu_\pi(s)\left(\frac{\ln T}{\theta}+\theta\sum_{k\in\mathcal{I}}\sum_{a\in\mathcal{A}}\pi_k(a|s)\left(\widehat{\beta}_k(s,a)+u_k(s,a)\right)^2\right)&&(\pi_k(\cdot|s)\in\bar{\Delta})\\&\le\sum_{s\in\mathcal{S}}\mu_\pi(s)\left(\frac{\ln T}{\theta}+4\theta|\mathcal{I}|H^2\iota^2\right)\le 4H\iota\sqrt{K\ln T}+4H\iota\ln T.&&(\text{Lemma 15 and definition of }\theta)\end{aligned}$$

Moreover, with probability at least $1-2\delta$,

$$\begin{aligned}&\sum_{k\in\mathcal{I}}J^{\pi_k,P_k,x_k}=\sum_{k\in\mathcal{I}}(J^{\pi_k,P_k,x_k}-J^{\pi_k,P,x_k})+\sum_{k\in\mathcal{I}}J^{\pi_k,P,x_k}\\&\le\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_{\pi_k}(s,a)[P_{k,s,a}-P_{s,a}]u'_k+\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_{\pi_k}(s,a)x_k(s,a)+K\epsilon_{\text{EVI}}&&(\text{Lemma 10})\\&\le\sum_{k\in\mathcal{I}}\sum_{s,a}\mu_{\pi_k}(s,a)\frac{3H\iota\sqrt{S\ln\frac{2SAT}{\delta}}}{\sqrt{N_k^+(s,a)}}+K\epsilon_{\text{EVI}}&&(\text{Lemma 9, definition of }\mathcal{Q}_k,\text{ and Lemma 14})\\&\le 24\iota\sqrt{S^2AH|\mathcal{I}|\ln\frac{2SAT}{\delta}}+60S^{1.5}AH\iota\ln^{3/2}\left(\frac{4SAT}{\delta}\right)+K\epsilon_{\text{EVI}}.&&(\text{Lemma 17})\end{aligned}$$

Substituting these back and by the definition of λ completes the proof. □

B.5. Proof of Theorem 1

Proof. For constraint violation, note that:

$$\sum_{t=1}^{T}(c(s_t,a_t)-\tau)=\sum_{k=1}^{K}\sum_{h=1}^{H}(c(s_h^k,a_h^k)-J^{\pi_k,c})+H\sum_{k=1}^{K}(J^{\pi_k,c}-\tau).$$

For the first term, by Lemma 18, we have $\sum_{k=1}^{K}\sum_{h=1}^{H}(c(s_h^k,a_h^k)-J^{\pi_k,c})\le\lambda+\tilde{\mathcal{O}}\left(t_{\text{mix}}\right)$ with probability at least $1-2\delta$. For the second term, by Lemma 19, we have $\lambda_{k+1}\ge\lambda_k+\widehat{J}_k+\epsilon-\tau$ with probability at least $1-6\delta$, and with probability

at least $1-\delta$,

$$\begin{aligned}
H\sum_{k=1}^{K}(J^{\pi_k,c}-\tau) &\le H\sum_{k=1}^{K}(J^{\pi_k,c}-\mathbb{E}_k[\widehat{J}_k])+H\sum_{k=1}^{K}(\mathbb{E}_k[\widehat{J}_k]-\widehat{J}_k)+H\sum_{k=1}^{K}(\lambda_{k+1}-\lambda_k-\epsilon)\\
&\le \frac{1}{T}+H\sqrt{2K\ln\frac{4T^3}{\delta}}+H(\lambda-K\epsilon) && \text{(Lemma 7 and Lemma 28)}\\
&\le \tilde{\mathcal{O}}\left(1\right)+\lambda+H(\lambda-K\epsilon). && \text{(definition of } \lambda)
\end{aligned}$$

When $\epsilon = 3\lambda/K$, that is, $3\lambda/K \ge (\tau - c^0)/2$, we have

$$\sum_{t=1}^{T}(c(s_t,a_t)-\tau) \le 2\lambda+\tilde{\mathcal{O}}\left(t_{\text{mix}}\right)+H(\lambda-K\epsilon)=\tilde{\mathcal{O}}\left(t_{\text{mix}}\right).$$

Otherwise, $K \le \frac{6\lambda}{\tau-c^0}$, which gives $T=\tilde{\mathcal{O}}\left(\frac{N^2H^2S^3A+N^2H^3S}{(\tau-c^0)^2}+\frac{NH^2S^2A}{\tau-c^0}\right)$ and the constraint violation is of the same order by $C_T \le T$.

For regret, note that $\sum_{k=1}^{K}(J^{\star,\epsilon}-J^{\pi_k,r})=\sum_{k=1}^{K}(J^{\pi^{\star,\epsilon},r-\frac{\lambda_k}{\eta}c}-J^{\pi_k,r-\frac{\lambda_k}{\eta}c})+\sum_{k=1}^{K}\frac{\lambda_k}{\eta}(J^{\pi^{\star,\epsilon},c}-J^{\pi_k,c})$, and with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{k=1}^{K}\frac{\lambda_k}{\eta}(J^{\pi^{\star,\epsilon},c}-J^{\pi_k,c}) &\le \sum_{k=1}^{K}\frac{\lambda_k}{\eta}\left(\tau-\epsilon-J^{\pi_k,c}\right) && \text{(definition of } \pi^{\star,\epsilon})\\
&\le \sum_{k=1}^{K}\frac{\lambda_k}{\eta}\left(\tau-\epsilon-\widehat{J}_k\right)+\sum_{k=1}^{K}\frac{\lambda_k}{\eta}(\widehat{J}_k-\mathbb{E}_k[\widehat{J}_k])+\frac{\lambda}{\eta T} && \text{(Lemma 7)}\\
&\le \frac{1}{\eta}\sum_{k=1}^{K}\lambda_k(\lambda_k-\lambda_{k+1})+\frac{\tau^2K}{\eta}+\tilde{\mathcal{O}}\left(\frac{\lambda}{\eta}\sqrt{K}\right),
\end{aligned}$$

where in the last step we apply Azuma's inequality and the following argument: if $\lambda_{k+1}>0$, then $\tau-\epsilon-\widehat{J}_k=\lambda_k-\lambda_{k+1}$ by the definition of λ_k. Otherwise, $\lambda_k \le \tau-\epsilon-\widehat{J}_k<\tau$ and $\lambda_k(\tau-\epsilon-\widehat{J}_k)\le\tau^2$. Moreover, $\sum_{k=1}^{K}\lambda_k(\lambda_k-\lambda_{k+1})=\frac{1}{2}\sum_{k=1}^{K}\left(\lambda_k^2-\lambda_{k+1}^2+(\lambda_{k+1}-\lambda_k)^2\right)\le\frac{K}{2}$ by $\lambda_1=0$ and $|\lambda_k-\lambda_{k+1}|\le 1$. Therefore, by Lemma 2 and definition of λ and η, with probability at least $1-4\delta$,

$$\begin{aligned}
\sum_{k=1}^{K}(J^{\star,\epsilon}-J^{\pi_k,r}) &\le \sum_{k=1}^{K}(J^{\pi^{\star,\epsilon},r-\frac{\lambda_k}{\eta}c}-J^{\pi_k,r-\frac{\lambda_k}{\eta}c})+\tilde{\mathcal{O}}\left(\frac{K}{\eta}+\frac{\lambda}{\eta}\sqrt{K}\right)\\
&=\tilde{\mathcal{O}}\left(\frac{N}{\tau-c^0}\left(\sqrt{S^3AT}+\sqrt{SHT}+S^2AH\right)+\frac{1}{(\tau-c^0)^2}\right).
\end{aligned}$$

Thus, with probability at least $1-2\delta$,

$$\begin{aligned}
&\sum_{t=1}^{T}(J^\star-r(s_t,a_t))=H\sum_{k=1}^{K}(J^\star-J^{\star,\epsilon})+H\sum_{k=1}^{K}(J^{\star,\epsilon}-J^{\pi_k,r})+\sum_{k=1}^{K}\sum_{h=1}^{H}(J^{\pi_k,r}-r(s_h^k,a_h^k))\\
&\le \frac{T\epsilon}{\tau-c^0}+\tilde{\mathcal{O}}\left(\frac{NH}{\tau-c^0}\left(\sqrt{S^3AT}+\sqrt{SHT}+S^2AH\right)+\frac{H}{(\tau-c^0)^2}\right) && \text{(Lemma 1 and Lemma 18)}\\
&=\tilde{\mathcal{O}}\left(\frac{NH}{\tau-c^0}\left(\sqrt{S^3AT}+\sqrt{SHT}+S^2AH\right)+\frac{H}{(\tau-c^0)^2}\right). && \text{(by the definition of } \epsilon)
\end{aligned}$$

Plugging in the definition of N and H completes the proof. □

B.6. Transition Estimation and Computation of u_k, P_k

Define $\mathcal{Q}_k$ as a transition confidence set based on Weissman's inequality (Lemma 30), such that $\mathcal{Q}_k=\cap_{s,a}\mathcal{Q}_{k,s,a}$, and

$$\mathcal{Q}_{k,s,a}=\left\{P':\left\|P'_{s,a}-\bar{P}_{k,s,a}\right\|_1\le\sqrt{\frac{S\ln\frac{2SAT}{\delta}}{N_k^+(s,a)}}\right\}.$$

We first show that P falls in $\mathcal{Q}_k$ with high probability.

Lemma 9. *With probability at least* $1-\delta$, $P \in \mathcal{Q}_k$ *for all* k.

Proof. For any (s,a), $n \in [T]$ and $m = S$, Lemma 30 gives with probability at least $1-\frac{\delta}{SAT}$: $\left\|P_{s,a} - \bar{P}^n_{s,a}\right\| \le \sqrt{\frac{S \ln \frac{2SAT}{\delta}}{n}}$, where $\bar{P}^n_{s,a}$ is the empirical distribution computed by n i.i.d samples from $P_{s,a}$. Taking a union bound over $(s,a) \in \mathcal{S} \times \mathcal{A}$, $n \in [T]$ proves the statement. □

Next, we show the computation procedure of u_k and P_k for a fixed episode k. Note that P_k is an approximation of $P_k^\star = \operatorname{argmax}_{P' \in \mathcal{Q}_k} J^{\pi_k, P', x_k}$, and finding $P_k^\star$ is equivalent to computing the optimal policy in an extended MDP $\widetilde{\mathcal{M}}_k$ with state space $\mathcal{S}$ and extended action space $\mathcal{Q}_k$, such that for any extended action $P' \in \mathcal{Q}_k$, the reward at (s, P') is $\sum_{a \in \mathcal{A}} \pi_k(a|s) r(s,a)$ and the transition probability to s' is $\sum_a \pi_k(a|s) P'_{s,a}(s')$. Note that since $\mathcal{Q}_k = \bigcap_{s,a} \mathcal{Q}_{k,s,a}$ where $\mathcal{Q}_{k,s,a}$ only puts constraints on transition at (s,a), any deterministic policy in $\widetilde{\mathcal{M}}_k$ can also be represented by an element in $\mathcal{Q}_k$. We adopt a variant of Extended Value Iteration (EVI) in (Jaksch et al., 2010, Theorem 7) to approximate $P_k^\star$, where we execute the following value iteration procedure in $\widetilde{\mathcal{M}}_k$,

$$u^0(s) = 0, \quad u^{i+1}(s) = \sum_{a \in \mathcal{A}} \pi_k(a|s) \left(x_k(s,a) + \max_{P \in \mathcal{Q}_k} P_{s,a} u^i \right). \tag{9}$$

We stop the iteration above at index $i^\star$, which is the first index i such that $\operatorname{sp}(u^{i+1} - u^i) \le \epsilon_{\text{EVI}} = \frac{1}{T}$. Then we define $u_k(s) = u^{i^\star+1}(s) - \min_{s'} u^{i^\star+1}(s')$, $u'_k(s) = u^{i^\star}(s) - \min_{s'} u^{i^\star}(s')$, $P_{k,s,a} = \operatorname{argmax}_{P \in \mathcal{Q}_k} P_{s,a} u^{i^\star}$ as the transition in u_k, and $u_k(s,a) = x_k(s,a) - \min_{s'} u^{i^\star+1}(s') + P_{k,s,a} u^{i^\star}$ so that $u_k(s) = \sum_a \pi_k(a|s) u_k(s,a)$, which is the function used in Algorithm 1. Also note that the maximization in Eq. (9) can be solved by (Jaksch et al., 2010, Figure 2).

Now we show that the value iteration in Eq. (9) always converges (specifically, the transition converges to $P_k^\star$) similar to (Jaksch et al., 2010, Theorem 7). First note that $\widetilde{\mathcal{M}}_k$ is communicating since $P \in \mathcal{Q}_k$ whose corresponding MDP is ergodic. Moreover, the transition chosen in each iteration of Eq. (9) is aperiodic and unichain. This is because in each iteration of Eq. (9) there is a "best" state and every state has non-zero probability transiting to the "best" state. Following the proof of (Jaksch et al., 2010, Theorem 7), we conclude that EVI in Eq. (9) converges.

Since P_k is aperiodic and unichain by the arguments above, there exist constant J^{π_k, P_k, x_k} such that $J^{\pi_k, P_k, x_k}(s) = J^{\pi_k, P_k, x_k}$. Importantly, by (Puterman, 1994, Theorem 8.5.6), we have:

$$|u^{i^\star+1}(s) - u^{i^\star}(s) - J^{\pi_k, P_k, x_k}| \le \epsilon_{\text{EVI}}. \tag{10}$$

This leads to the following approximated value difference lemma.

Lemma 10. $J^{\pi, P, x_k} - J^{\pi_k, P_k, x_k} = \sum_{s,a} \mu_\pi(s) \left(\pi(a|s) - \pi_k(a|s)\right) u_k(s,a) + \sum_{s,a} \mu_\pi(s,a)(P_{s,a} - P_{k,s,a}) u'_k + \delta_{\text{EVI}}$, *where* $|\delta_{\text{EVI}}| \le \epsilon_{\text{EVI}}$.

Proof. Define $u^i(s,a) = x_k(s,a) + \max_{P \in \mathcal{Q}_k} P_{s,a} u^{i-1}$ so that $u^i(s) = \sum_a \pi_k(a|s) u^i(s,a)$. Since P is ergodic, we have:

$$\begin{aligned}
J^{\pi, P, x_k} &= \sum_{s,a} \mu_\pi(s,a) x_k(s,a) = \sum_{s,a} \mu_\pi(s,a) \left(u^{i^\star+1}(s,a) - P_{k,s,a} u^{i^\star} \right) \\
&= \sum_{s,a} \mu_\pi(s,a) \left(u^{i^\star+1}(s,a) - u^{i^\star+1}(s) \right) + \sum_{s,a} \mu_\pi(s,a) \left(u^{i^\star+1}(s) - P_{s,a} u^{i^\star} \right) + \sum_{s,a} \mu_\pi(s,a)(P_{s,a} - P_{k,s,a}) u^{i^\star} \\
&= \sum_{s,a} \mu_\pi(s,a) \left(u_k(s,a) - u_k(s) \right) + \sum_{s,a} \mu_\pi(s,a) \left(u^{i^\star+1}(s) - P_{s,a} u^{i^\star} \right) + \sum_{s,a} \mu_\pi(s,a)(P_{s,a} - P_{k,s,a}) u'_k.
\end{aligned}$$

Let $\delta_{\text{EVI}} = \sum_{s,a} \mu_\pi(s,a) \left(u^{i^\star+1}(s) - P_{s,a} u^{i^\star} \right)$. By Eq. (10), we have

$$\delta_{\text{EVI}} \le \sum_{s,a} \mu_\pi(s,a) \left(u^{i^\star}(s) - P_{s,a} u^{i^\star} + J^{\pi_k, P_k, x_k} + \epsilon_{\text{EVI}} \right) = J^{\pi_k, P_k, x_k} + \epsilon_{\text{EVI}}. \quad (\mu_\pi(s') = \textstyle\sum_{s,a} \mu_\pi(s,a) P_{s,a}(s'))$$

Showing $\delta_{\text{EVI}} \ge -\epsilon_{\text{EVI}}$ is similar. Further by $\sum_a \mu_\pi(s,a)(u_k(s,a) - u_k(s)) = \mu_\pi(s) \sum_a (\pi(a|s) - \pi_k(a|s)) u_k(s,a)$, the statement is proved. □

B.7. Auxiliary Lemmas

Lemma 11. *(Wei et al., 2020, Corollary 13.1) For any ergodic MDP with mixing time t_{mix}, we have $\left\|(P^\pi)^t_{s,\cdot} - \mu_\pi\right\|_1 \le 2\cdot 2^{-t/t_{mix}}$ for all policy π, state $s\in\mathcal{S}$, and $t \ge 2t_{mix}$.*

Lemma 12. *(Wei et al., 2020, Corollary 13.2) Let $N = 4t_{mix}\log_2 T$. For an ergodic MDP with mixing time $t_{mix} < T/4$, we have for all π: $\sum_{t=N}^{\infty}\left\|(P^\pi)^t_{s,\cdot} - \mu_\pi\right\|_1 \le \frac{1}{T^3}$.*

Lemma 13. *(Wei et al., 2020, Lemma 14) For an ergodic MDP with mixing time t_{mix}, utility function $d\in[0,1]^{\mathcal{S}\times\mathcal{A}}$, and any π, s, a, $|v^{\pi,d}(s)| \le 5t_{mix}$ and $|q^{\pi,d}(s,a)| \le 6t_{mix}$.*

Lemma 14. *Under the event of Lemma 9, $\max\{\mathrm{sp}(u_k), \mathrm{sp}(u'_k)\} \le 4t_{mix}t_{hit}\lceil\log_2(4t_{hit})\rceil\iota \le \frac{H\iota}{4}$.*

Proof. For a fixed k, it suffices to show that for any two states s, s' and $H\ge 1$, $u^H(s) - u^H(s') = \tilde{\mathcal{O}}((\lambda/\eta)t_{\text{hit}}t_{\text{mix}})$, where u^i defined in Eq. (9) is the optimal value function of taking H steps in $\widetilde{\mathcal{M}}_k$. Without loss of generality, assume $u^H(s)\ge u^H(s')$. Define random variable τ as the number of steps it takes to transits from state s' to s. Then by Lemma 9, $u^H(s') \ge \mathbb{E}_\tau[u^{H-\min\{H,\tau\}}(s)|\pi_k, P]$ (the right hand side is a lower bound of the expected reward of a history-dependent policy in $\widetilde{\mathcal{M}}_k$ which follows P at first and then switches to P_k when reaching s, and it is dominated by $u_H(s')$ by the Markov property). Thus by $x_k(s,a)\le 2\iota$,

$$\begin{aligned} u^H(s) - u^H(s') &\le u^H(s) - \mathbb{E}_\tau[u^{H-\min\{H,\tau\}}(s)|\pi_k, P] = \mathbb{E}_\tau[u^H(s) - u^{H-\min\{H,\tau\}}(s)|\pi_k, P] \\ &\le 2\mathbb{E}_\tau[\tau|\pi_k, P]\iota \le 4t_{\text{mix}}t_{\text{hit}}\lceil\log_2(4t_{\text{hit}})\rceil\iota. \end{aligned}$$

For the last inequality above, note that when $t = t_{\text{mix}}\lceil\log_2(4t_{\text{hit}})\rceil$, we have $\left\|(P^{\pi_k})^t_{s,\cdot} - \mu_{\pi_k}\right\|_\infty \le \frac{1}{2t_{\text{hit}}}$ for any $s\in\mathcal{S}$ by Lemma 11. Therefore, $(P^{\pi_k})^t_{s,s'} \ge \frac{1}{2}\mu_{\pi_k}(s') \ge \frac{1}{2t_{\text{hit}}}$ for any $s, s'\in\mathcal{S}$. This implies that we can reach any state at least once by taking $2t\cdot t_{\text{hit}}$ steps in expectation, that is, $\mathbb{E}_\tau[\tau|\pi_k, P] \le 2t\cdot t_{\text{hit}}$. The second inequality in the statement follows directly from the definition of H. □

Lemma 15. *Under the event of Lemma 9, $|\widehat{\beta}_k(s,a) + u_k(s,a)| \le 2H\iota$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.*

Proof. Note that $\widehat{\beta}_k(s,a) \le \frac{2\lambda}{\eta}(N+1) \le H\iota$. Define $s^\star = \mathrm{argmin}_s u^{i^\star+1}(s)$. By Lemma 14,

$$\begin{aligned} u_k(s,a) &= x_k(s,a) + P_{k,s,a}u^{i^\star} - u^{i^\star+1}(s^\star) = x_k(s,a) - \sum_{a'}\pi_k(a'|s^\star)x_k(s^\star,a') + (P_{k,s,a} - (P_k^{\pi_k})_{s^\star,\cdot})u^{i^\star} \\ &\le 2\iota + \mathrm{sp}(u^{i^\star}) \le H\iota. \end{aligned}$$

This completes the proof. □

Lemma 16. *(Wei et al., 2020, Lemma 15) For any two policies π, π' and utility function d,*

$$J^{\pi,d} - J^{\pi',d} = \sum_{s,a}\mu_\pi(s)(\pi(a|s) - \pi'(a|s))q^{\pi',d}(s,a).$$

Lemma 17. *With probability at least $1-\delta$, for any $\mathcal{I}\subseteq[K]$, $\sum_{k\in\mathcal{I}}\sum_{s,a}\frac{\mu_{\pi_k}(s,a)}{\sqrt{N_k^+(s,a)}} = 8\sqrt{SA|\mathcal{I}|/H} + 20SA\ln\frac{4T}{\delta}$.*

Proof. Define $n_k(s,a) = \sum_{h=N+1}^{H}\mathbb{I}\{s_h^k = s, a_h^k = a\}$. Note that $\mathbb{E}_k[n_k(s,a)] = \sum_{h=N}^{H-1}(P^{\pi_k})^h_{s_1^k,s}\pi_k(a|s)$. Moreover, by

Lemma 11, for any state s' and $h \geq N$: $\mu_{\pi_k}(s,a) - (P^{\pi_k})^h_{s',s}\pi_k(a|s) \leq 1/T^2$. Therefore,

$$\begin{aligned}
\sum_{k\in\mathcal{I}}\sum_{s,a}\frac{\mu_{\pi_k}(s,a)}{\sqrt{N_k^+(s,a)}} &\leq \sum_{k\in\mathcal{I}}\sum_{s,a}\frac{1}{T^2} + \frac{1}{H-N}\sum_{k\in\mathcal{I}}\sum_{s,a}\sum_{h=N}^{H-1}\frac{(P^{\pi_k})^h_{s_1^k,s}\pi_k(a|s)}{\sqrt{N_k^+(s,a)}} \\
&\leq SA/(HT) + \frac{2}{H-N}\sum_{k\in\mathcal{I}}\sum_{s,a}\frac{n_k(s,a)}{\sqrt{N_k^+(s,a)}} + 16\ln\frac{4T}{\delta} \\
&\qquad \text{(Lemma 32 with a union bound over } T \text{ possible values of } \mathcal{I}_{[1]}) \\
&\leq \frac{2}{H-N}\sum_{k\in\mathcal{I}}\sum_{s,a}\frac{n_k(s,a)}{\sqrt{N_{k+1}^+(s,a)}} + 2\sum_{k\in\mathcal{I}}\sum_{s,a}\left(\frac{1}{\sqrt{N_k^+(s,a)}} - \frac{1}{\sqrt{N_{k+1}^+(s,a)}}\right) + 17SA\ln\frac{4T}{\delta} \\
&\leq \frac{8}{H}\sqrt{SAH|\mathcal{I}|} + 20SA\ln\frac{4T}{\delta} = 8\sqrt{SA|\mathcal{I}|/H} + 20SA\ln\frac{4T}{\delta},
\end{aligned}$$

where the last inequality is by $\sum_{k\in\mathcal{I}}\frac{n_k(s,a)}{\sqrt{N_{k+1}^+(s,a)}} \leq 2\sqrt{\sum_{k\in\mathcal{I}} n_k(s,a)}$, Cauchy Schwarz inequality, and $H - N \geq \frac{H}{2}$. □

Lemma 18. *For any utility function $d \in [0,1]^{\mathcal{S}\times\mathcal{A}}$, we have $\left|\sum_{k=1}^K\sum_{h=1}^H d(s_h^k,a_h^k) - J^{\pi_k,d}\right| \leq 12t_{mix}\sqrt{2T\ln\frac{4T^3}{\delta}} + 33N^2\sqrt{K\ln T} + \tilde{\mathcal{O}}(t_{mix}) \leq \lambda + \tilde{\mathcal{O}}(t_{mix})$ with probability at least $1-2\delta$.*

Proof. With probability at least $1-2\delta$,

$$\begin{aligned}
&\left|\sum_{k=1}^K\sum_{h=1}^H (d(s_h^k,a_h^k) - J^{\pi_k,d})\right| = \left|\sum_{k=1}^K\sum_{h=1}^H \left(q^{\pi_k,d}(s_h^k,a_h^k) - P_h^k v^{\pi_k,d}\right)\right| \qquad \text{(Eq. (2))} \\
&= \left|\sum_{k=1}^K\sum_{h=1}^H \left(q^{\pi_k,d}(s_h^k,a_h^k) - v^{\pi_k,d}(s_h^k)\right) + \sum_{k=1}^K (v^{\pi_k,d}(s_1^k) - v^{\pi_k,d}(s_{H+1}^k)) + \sum_{k=1}^K\sum_{h=1}^H (v^{\pi_k,d}(s_{h+1}^k) - P_h^k v^{\pi_k,d})\right| \\
&\qquad (\textstyle\sum_{h=1}^H (v^{\pi_k,d}(s_h^k) - v^{\pi_k,d}(s_{h+1}^k)) = v^{\pi_k,d}(s_1^k) - v^{\pi_k,d}(s_{H+1}^k)) \\
&\leq \left|\sum_{k=1}^K\sum_{h=1}^H \left(q^{\pi_k,d}(s_h^k,a_h^k) - v^{\pi_k,d}(s_h^k)\right) + \sum_{k=2}^K (v^{\pi_k,d}(s_1^k) - v^{\pi_{k-1},d}(s_1^k)) + \sum_{k=1}^K\sum_{h=1}^H (v^{\pi_k,d}(s_{h+1}^k) - P_h^k v^{\pi_k,d})\right| + \tilde{\mathcal{O}}(t_{\text{mix}}) \\
&\qquad (s_{H+1}^k = s_1^{k+1} \text{ and Lemma 13}) \\
&\leq 12t_{\text{mix}}\sqrt{2T\ln\frac{4T^3}{\delta}} + 65\theta HN^2K\iota + \tilde{\mathcal{O}}(t_{\text{mix}}) \leq 12t_{\text{mix}}\sqrt{2T\ln\frac{4T^3}{\delta}} + 33N^2\sqrt{K\ln T} + \tilde{\mathcal{O}}(t_{\text{mix}}). \\
&\qquad \text{(Lemma 3, Lemma 13, and Lemma 28)}
\end{aligned}$$

The second inequality directly follows from the definition of λ. □

Lemma 19. *With probability at least $1-6\delta$, $\lambda_k < \lambda$ for any k, that is, the upper bound truncation of λ_k is never triggered.*

Proof. We prove this by induction on k. The base case $k=1$ is clearly true. For $k>1$, if $\lambda_k \leq \frac{2(\eta+1)}{\tau-c^0}$, the statement is proved. Otherwise, let $j = \max\{j' < k : \lambda_{j'} \leq \frac{2(\eta+1)}{\tau-c^0}, \lambda_{j'+1} > \frac{2(\eta+1)}{\tau-c^0}\}$. We have:

$$\lambda_k^2 = \lambda_j^2 + \sum_{i=j}^{k-1}(\lambda_{i+1}^2 - \lambda_i^2) \leq \left(\frac{2(\eta+1)}{\tau-c^0}\right)^2 + \sum_{i=j}^{k-1}\left(2\lambda_i(\lambda_{i+1}-\lambda_i) + (\lambda_{i+1}-\lambda_i)^2\right).$$

Note that $\lambda_i > 0$ for $j < i \le k$. Therefore, with probability at least $1-6\delta$,

$$\sum_{i=j}^{k-1}\left(2\lambda_i(\lambda_{i+1}-\lambda_i)+(\lambda_{i+1}-\lambda_i)^2\right) \le \sum_{i=j}^{k-1}\left(2\lambda_i(\widehat{J}_i+\epsilon-\tau)+1\right) \qquad \text{(definition of } \lambda_i \text{ and } |\lambda_{i+1}-\lambda_i|\le 1)$$

$$\le \sum_{i=j}^{k-1}\left(2\lambda_i(2\mathbb{E}_i[\widehat{J}_i]+\epsilon-\tau)+1\right)+32\lambda\ln\frac{4T}{\delta} \le \sum_{i=j}^{k-1}\left(4\lambda_i(J^{\pi_i,c}+\epsilon-\tau)+1\right)+33\lambda\ln\frac{4T}{\delta}$$

(Lemma 32 with a union bound over T possible values of j, $\lambda_i \le \lambda$ by definition, and Lemma 7)

$$\le 4\sum_{i=j}^{k-1}\left(\lambda_i(J^{\pi_i,c}+\epsilon-\tau)+1\right)+33\lambda\ln\frac{4T}{\delta} = 4\sum_{i=j}^{k-1}\left(\eta J^{\pi_i,r}-\eta J^{\pi_i,r-\frac{\lambda_i}{\eta}c}-\lambda_i(\tau-\epsilon)+1\right)+33\lambda\ln\frac{4T}{\delta}$$

$$\le 4\sum_{i=j}^{k-1}\left(\eta J^{\pi_i,r}-\eta J^{\pi^0,r-\frac{\lambda_i}{\eta}c}-\lambda_i(\tau-\epsilon)+1\right)+33\lambda\ln\frac{4T}{\delta}+\frac{\eta\lambda}{\tau-c^0} \qquad \text{(Lemma 2)}$$

$$\le 4\sum_{i=j}^{k-1}\left(\eta-\frac{\tau-c^0}{2}\lambda_i+1\right)+33\lambda\ln\frac{4T}{\delta}+\frac{\eta\lambda}{\tau-c^0} \qquad (J^{\pi_i,r}-J^{\pi^0,r}\le 1,\ J^{\pi^0,c}=c^0, \text{ and } \epsilon\le\tfrac{\tau-c^0}{2})$$

$$\le 4(\eta+1)+33\lambda\ln\frac{4T}{\delta}+\frac{\eta\lambda}{\tau-c^0}. \qquad (\lambda_i > \tfrac{2(\eta+1)}{\tau-c^0} \text{ for } j<i\le k)$$

Then by $\frac{\lambda}{4} > \frac{4(\eta+1)}{\tau-c^0}$ and $\eta \ge 132(\tau-c^0)\ln\frac{4T}{\delta}$, we have $\lambda_k = \lambda_k^2/\lambda_k \le \frac{2(\eta+1)}{\tau-c^0}+2+\frac{\lambda}{8}+\frac{\lambda}{2} < \lambda$. □

C. Omitted Details for Section 4

As a standard practice, we first show that the true transition lies in the transition confidence sets with high probability, and provide some key lemmas related to transition estimation.

Lemma 20. *With probability at least* $1-\delta$, $\widetilde{P} \in \mathcal{P}_k, \forall k$.

Proof. For any $(s,a)\in\mathcal{S}\times\mathcal{A}, s'\in\mathcal{S}$, by Lemma 31 and $N_{K+1}(s,a)\le T$, we have with probability at least $1-\frac{\delta}{S^2A}$,

$$\left|P_{s,a}(s')-\bar{P}_{k,s,a}(s')\right| \le 4\sqrt{\bar{P}_{k,s,a}(s')\alpha_k(s,a)}+28\alpha_k(s,a).$$

By a union bound over all $(s,a)\in\mathcal{S}\times\mathcal{A}$, $s'\in\mathcal{S}$ and $\widetilde{P}_{s,a,h}=P_{s,a}$, the statement is proved. □

Lemma 21. *Under the event of Lemma 20,* $\left|P'_{s,a,h}(s')-P_{s,a}(s')\right| \le 8\sqrt{P_{s,a}(s')\alpha_k(s,a)}+136\alpha_k(s,a) \triangleq \epsilon_k^\star(s,a,s')$ *for any* $P'\in\mathcal{P}_k$.

Proof. By $\widetilde{P}\in\mathcal{P}_k$, we have for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, and $s'\in\mathcal{S}$:

$$\bar{P}_{k,s,a}(s') \le P_{s,a}(s')+4\sqrt{\bar{P}_{k,s,a}(s')\alpha_k(s,a)}+28\alpha_k(s,a).$$

Applying $x^2\le ax+b \implies x\le a+\sqrt{b}$ with $a=4\sqrt{\alpha_k(s,a)}$ and $b=P_{s,a}(s')+28\alpha_k(s,a)$, we have

$$\sqrt{\bar{P}_{k,s,a}(s')} \le 4\sqrt{\alpha_k(s,a)}+\sqrt{P_{s,a}(s')+28\alpha_k(s,a)} \le \sqrt{P_{s,a}(s')}+10\sqrt{\alpha_k(s,a)}.$$

Substituting this back to right-hand side of the inequality in Eq. (7), we have

$$4\sqrt{\bar{P}_{k,s,a}(s')\alpha_k(s,a)}+28\alpha_k(s,a) \le 4\sqrt{P_{s,a}(s')\alpha_k(s,a)}+68\alpha_k(s,a).$$

By $\widetilde{P}, P'\in\mathcal{P}_k$, Eq. (7), and the triangle inequality $|P'_{s,a,h}(s')-P_{s,a}(s')| \le |P'_{s,a,h}(s')-\bar{P}_{k,s,a}(s')|+|\bar{P}_{k,s,a}(s')-P_{s,a}(s')|$, the statement is proved. □

Lemma 22. *Under the event of Lemma 20, Algorithm 3 ensuers $\nu_{\widetilde{\pi}^\star,\widetilde{P},s_1^k}$ lies in the domain of Eq. (6).*

Proof. By Lemma 4, we have:

$$\left\langle \nu_{\widetilde{\pi}^\star,\widetilde{P},s_1^k}, c \right\rangle = V_1^{\widetilde{\pi}^\star,c}(s_1^k) \le \mathrm{sp}_c^\star + HJ^{\pi^\star,c} \le \mathrm{sp}_c^\star + H\tau. \tag{11}$$

Then by $\widetilde{P} \in \mathcal{P}_k$, the statement is proved. □

Lemma 23. *Under the event of Lemma 20, Algorithm 4 ensures $\nu_{\widetilde{\pi}^\star,\widetilde{P},s_1^k}$ lies in the domain of Eq. (8).*

Proof. Note that Eq. (11) still holds. Moreover, by Lemma 4, for any two states s, s' and $h \in [H]$:

$$|V_h^{\widetilde{\pi}^\star}(s) - V_h^{\widetilde{\pi}^\star}(s')| \le |V_h^{\widetilde{\pi}^\star}(s) - (H-h+1)J^{\pi^\star,r}| + |V_h^{\widetilde{\pi}^\star}(s') - (H-h+1)J^{\pi^\star,r}| \le 2\mathrm{sp}_r^\star,$$
$$|V_h^{\widetilde{\pi}^\star,c}(s) - V_h^{\widetilde{\pi}^\star,c}(s')| \le |V_h^{\widetilde{\pi}^\star,c}(s) - (H-h+1)J^{\pi^\star,c}| + |V_h^{\widetilde{\pi}^\star,c}(s') - (H-h+1)J^{\pi^\star,c}| \le 2\mathrm{sp}_c^\star.$$

Then by $\widetilde{P} \in \mathcal{P}_k$, the statement is proved. □

C.1. Proof of Lemma 4

Proof. For any state s and $h \in [H]$, we have:

$$\begin{aligned}
V_h^{\widetilde{\pi},d}(s) - (H-h+1)J^{\pi,d} &= \mathbb{E}\left[\sum_{h'=h}^{H}(d(s_{h'},a_{h'}) - J^{\pi,d})\middle|\widetilde{\pi},\widetilde{P},s_h=s\right] \\
&= \mathbb{E}\left[\sum_{h'=h}^{H}(q^{\pi,d}(s_{h'},a_{h'}) - P_{s_{h'},a_{h'}}v^{\pi,d})\middle|\widetilde{\pi},\widetilde{P},s_h=s\right] && \text{(Eq. (2))} \\
&= \mathbb{E}\left[\sum_{h'=h}^{H}(v^{\pi,d}(s_{h'}) - v^{\pi,d}(s_{h'+1}))\middle|\widetilde{\pi},\widetilde{P},s_h=s\right] && \text{(definition of } \widetilde{\pi} \text{ and } \widetilde{P}) \\
&= v^{\pi,d}(s) - \mathbb{E}\left[v^{\pi,d}(s_{H+1})\middle|\widetilde{\pi},\widetilde{P},s_h=s\right].
\end{aligned}$$

Thus, $|V_h^{\widetilde{\pi},d}(s) - (H-h+1)J^{\pi,d}| \le \mathrm{sp}(v^{\pi,d})$ and the statement is proved. □

C.2. Proof of Lemma 5

Proof. We condition on the event of Lemma 20, which happens with probability at least $1-\delta$. Note that with probability at least $1-\delta$:

$$\begin{aligned}
\left|\sum_{k=1}^{K}(V_1^{\pi_k,P_k,d}(s_1^k) - V_1^{\pi_k,d}(s_1^k))\right| &= \left|\sum_{k=1}^{K}\mathbb{E}\left[\sum_{h=1}^{H}(P_{k,s_h^k,a_h^k,h} - P_h^k)V_{h+1}^{\pi_k,P_k,d}\middle|\pi_k,P\right]\right| && \text{(Lemma 26)} \\
&\le \sum_{k=1}^{K}\mathbb{E}\left[\sum_{h=1}^{H}\left|(P_{k,s_h^k,a_h^k,h} - P_h^k)V_{h+1}^{\pi_k,P_k,d}\right|\middle|\pi_k,P\right] && \text{(Jensen's inequality)} \\
&\le 2\sum_{k=1}^{K}\sum_{h=1}^{H}\left|(P_{k,s_h^k,a_h^k,h} - P_h^k)V_{h+1}^{\pi_k,P_k,d}\right| + \tilde{\mathcal{O}}\left(H^2\right) && \text{(Lemma 32)} \\
&= \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_h^k, V_{h+1}^{\pi_k,P_k,d})} + H^2S^2A\right). && \text{(Lemma 25)}
\end{aligned}$$

Then by Lemma 24 and $H = (T/S^2A)^{1/3}$, with probability at least $1-2\delta$,

$$\left|\sum_{k=1}^{K}V_1^{\pi_k,d}(s_1^k) - V_1^{\pi_k,P_k,d}(s_1^k)\right| = \tilde{\mathcal{O}}\left(\sqrt{S^2A(H^2K + H^3S^2A)} + H^2S^2A\right) = \tilde{\mathcal{O}}\left(\sqrt{S^2AH^2K} + H^2S^2A\right).$$

This completes the proof. □

C.3. Proof of Lemma 6

Proof. Define $\bar{V}_h^{\pi_k,P_k,d}(s) = V_h^{\pi_k,P_k,d}(s) - \min_{s'} V_h^{\pi_k,P_k,d}(s')$ and $\bar{Q}_h^{\pi_k,P_k,d}(s,a) = Q_h^{\pi_k,P_k,d}(s,a) - \min_{s'} V_h^{\pi_k,P_k,d}(s')$ so that $\bar{V}_h^{\pi_k,P_k,d}(s) \in [0,B]$ and

$$\begin{aligned}
&\left|\bar{Q}_h^{\pi_k,P_k,d}(s,a)\right| = \left|Q_h^{\pi_k,P_k,d}(s,a) - V_h^{\pi_k,P_k,d}(s^\star)\right| && (s^\star = \operatorname{argmin}_s V_h^{\pi_k,P_k,d}(s)) \\
&\le \left|d(s,a) - \sum_{a\in\mathcal{A}} \pi_k(a|s^\star) d(s^\star,a)\right| + \left|P_{k,s,a} V_{h+1}^{\pi_k,P_k,d} - (P_k^{\pi_k})_{s,\cdot} V_{h+1}^{\pi_k,P_k,d}\right| \le B+1.
\end{aligned}$$

Also define $\mathbb{I}_s(s') = \mathbb{I}\{s = s'\}$. Then with probability at least $1-2\delta$,

$$\begin{aligned}
&\sum_{k=1}^K \left(V_1^{\pi_k,P_k,d}(s_1^k) - \sum_{h=1}^H d(s_h^k,a_h^k)\right) \\
&= \sum_{k=1}^K \left(V_1^{\pi_k,P_k,d}(s_1^k) - Q_1^{\pi_k,P_k,d}(s_1^k,a_1^k) + Q_1^{\pi_k,P_k,d}(s_1^k,a_1^k) - d(s_1^k,a_1^k) - \sum_{h=2}^H d(s_h^k,a_h^k)\right) \\
&= \sum_{k=1}^K \left(V_1^{\pi_k,P_k,d}(s_1^k) - Q_1^{\pi_k,P_k,d}(s_1^k,a_1^k) + (P_{k,s_1^k,a_1^k,1} - P_1^k) V_2^{\pi_k,P_k,d} + (P_1^k - \mathbb{I}_{s_2^k}) V_2^{\pi_k,P_k,d}\right) \\
&\quad + \sum_{k=1}^K \left(V_2^{\pi_k,P_k,d}(s_2^k) - \sum_{h=2}^H d(s_h^k,a_h^k)\right) \\
&= \sum_{k=1}^K \sum_{h=1}^H \left(V_h^{\pi_k,P_k,d}(s_h^k) - Q_h^{\pi_k,P_k,d}(s_h^k,a_h^k) + (P_{k,s_h^k,a_h^k,h} - P_h^k) V_{h+1}^{\pi_k,P_k,d} + (P_h^k - \mathbb{I}_{s_{h+1}^k}) V_{h+1}^{\pi_k,P_k,d}\right) \\
&\qquad\qquad \text{(repeat the decomposition above)} \\
&= \sum_{k=1}^K \sum_{h=1}^H \left(\bar{V}_h^{\pi_k,P_k,d}(s_h^k) - \bar{Q}_h^{\pi_k,P_k,d}(s_h^k,a_h^k) + (P_{k,s_h^k,a_h^k,h} - P_h^k) \bar{V}_{h+1}^{\pi_k,P_k,d} + (P_h^k - \mathbb{I}_{s_{h+1}^k}) \bar{V}_{h+1}^{\pi_k,P_k,d}\right) \\
&= \tilde{\mathcal{O}}\left((B+1)\sqrt{T} + BS\sqrt{AT} + BHS^2A\right). \qquad \text{(Lemma 28, Lemma 25, and } \mathbb{V}(P_h^k, \bar{V}_{h+1}^{\pi_k,P_k,d}) \le B^2)
\end{aligned}$$

□

C.4. Auxiliary Lemmas

Lemma 24. *Under the event of Lemma 20, for any utility function $d \in [0,1]^{\mathcal{S}\times\mathcal{A}}$, with probability at least $1-2\delta$, $\sum_{k=1}^K \sum_{h=1}^H \mathbb{V}(P_h^k, V_{h+1}^{\pi_k,P_k,d}) = \tilde{\mathcal{O}}(H^2(K+\sqrt{T}) + H^3S^2A)$.*

Proof. We decompose the variance into four terms:

$$\begin{aligned}
&\sum_{k=1}^K \sum_{h=1}^H \mathbb{V}(P_h^k, V_{h+1}^{\pi_k,P_k,d}) = \sum_{k=1}^K \sum_{h=1}^H \left(P_h^k (V_{h+1}^{\pi_k,P_k,d})^2 - (P_h^k V_{h+1}^{\pi_k,P_k,d})^2\right) \\
&= \sum_{k=1}^K \sum_{h=1}^H \left(P_h^k (V_{h+1}^{\pi_k,P_k,d})^2 - V_{h+1}^{\pi_k,P_k,d}(s_{h+1}^k)^2\right) + \sum_{k=1}^K \sum_{h=1}^H \left(V_{h+1}^{\pi_k,P_k,d}(s_{h+1}^k)^2 - V_h^{\pi_k,P_k,d}(s_h^k)^2\right) \\
&\quad + \sum_{k=1}^K \sum_{h=1}^H \left(V_h^{\pi_k,P_k,d}(s_h^k)^2 - Q_h^{\pi_k,P_k,d}(s_h^k,a_h^k)^2\right) + \sum_{k=1}^K \sum_{h=1}^H \left(Q_h^{\pi_k,P_k,d}(s_h^k,a_h^k)^2 - (P_h^k V_{h+1}^{\pi_k,P_k,d})^2\right).
\end{aligned}$$

For the first term, by Lemma 29, with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H} P_h^k(V_{h+1}^{\pi_k,P_k,d})^2 - V_{h+1}^{\pi_k,P_k,d}(s_{h+1}^k)^2 = \tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_h^k,(V_{h+1}^{\pi_k,P_k,d})^2)} + H^2\right)$$
$$= \tilde{\mathcal{O}}\left(H\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_h^k,V_{h+1}^{\pi_k,P_k,d})} + H^2\right). \qquad \text{(Lemma 27)}$$

The second term is upper bounded by 0 by $V_{H+1}^{\pi_k,P_k,d}(s)=0$ for $s\in\mathcal{S}$. For the third term, by Cauchy-Schwarz inequality and Lemma 28, with probability at least $1-\delta$:

$$\sum_{k=1}^{K}\sum_{h=1}^{H} V_h^{\pi_k,P_k,d}(s_h^k)^2 - Q_h^{\pi_k,P_k,d}(s_h^k,a_h^k)^2 \le \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\sum_a \pi_k(a|s_h^k,h)Q_h^{\pi_k,P_k,d}(s_h^k,a)^2 - Q_h^{\pi_k,P_k,d}(s_h^k,a_h^k)^2\right)$$
$$= \tilde{\mathcal{O}}\left(H^2\sqrt{T}\right).$$

For the fourth term, by $a^2-b^2=(a+b)(a-b)$ and $\left\|V_h^{\pi_k,P_k,d}\right\|_\infty, \left\|Q_h^{\pi_k,P_k,d}\right\|_\infty \le H$:

$$\sum_{k=1}^{K}\sum_{h=1}^{H} Q_h^{\pi_k,P_k,d}(s_h^k,a_h^k)^2 - (P_h^k V_{h+1}^{\pi_k,P_k,d})^2 \le 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\left|Q_h^{\pi_k,P_k,d}(s_h^k,a_h^k) - P_h^k V_{h+1}^{\pi_k,P_k,d}\right|$$
$$\le 2H^2K + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\left|(P_{k,s_h^k,a_h^k,h} - P_h^k)V_{h+1}^{\pi_k,P_k,d}\right| = \tilde{\mathcal{O}}\left(H^2K + H\sqrt{S^2A\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_h^k,V_{h+1}^{\pi_k,P_k,d})} + H^3S^2A\right).$$
(Lemma 25)

Putting everything together, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_h^k,V_{h+1}^{\pi_k,P_k,d})$$
$$= \tilde{\mathcal{O}}\left(H\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_h^k,V_{h+1}^{\pi_k,P_k,d})} + H^2(K+\sqrt{T}) + H\sqrt{S^2A\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_h^k,V_{h+1}^{\pi_k,P_k,d})} + H^3S^2A\right).$$

Solving a quadratic inequality, we get $\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_h^k,V_{h+1}^{\pi_k,P_k,d}) = \tilde{\mathcal{O}}(H^2(K+\sqrt{T})+H^3S^2A)$. □

Lemma 25. *Under the event of Lemma 20, for any value function V with $V_h\in[0,B]^{\mathcal{S}}, \forall h\in[H]$, we have:*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left|(P_{k,s_h^k,a_h^k,h} - P_h^k)V_{h+1}\right| = \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_h^k,V_{h+1})} + BHS^2A\right).$$

Proof. Define $\mathbb{I}_k = \mathbb{I}\{\forall(s,a): N_{k+1}(s,a)\le 2N_k(s,a)\}$ and $z_h^k(s') = V_{h+1}(s') - P_h^kV_{h+1}$. By Lemma 21,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left|(P_{k,s_h^k,a_h^k,h} - P_h^k)V_{h+1}\right| = \sum_{k=1}^{K}\sum_{h=1}^{H}\left|(P_{k,s_h^k,a_h^k,h} - P_h^k)z_h^k\right| \le \sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{B, \sum_{s'}\epsilon_k^\star(s_h^k,a_h^k,s')|z_h^k(s')|\right\}$$
$$\le 2\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{B, \sum_{s'}\epsilon_{k+1}^\star(s_h^k,a_h^k,s')|z_h^k(s')|\right\} + BH\sum_{k=1}^{K}\mathbb{I}_k^c.$$

Note that $\sum_{k=1}^{K}\mathbb{I}_k^c=\tilde{\mathcal{O}}(SA)$ by definition. Thus it suffices to bound $\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'}\epsilon_{k+1}^{\star}(s_h^k,a_h^k,s')|z_h^k(s')|$. Note that:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'}\epsilon_{k+1}^{\star}(s_h^k,a_h^k,s')|z_h^k(s')| = \tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'}\sqrt{\frac{P_h^k(s')z_h^k(s')^2}{N_{k+1}^+(s_h^k,a_h^k)}}+\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{SB}{N_{k+1}^+(s_h^k,a_h^k)}\right) && \text{(definition of } \epsilon_k^\star\text{)}\\
&=\tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{S\mathbb{V}(P_h^k,V_{h+1})}{N_{k+1}^+(s_h^k,a_h^k)}}+BS^2A\right)\\
&=\tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{S}{N_{k+1}^+(s_h^k,a_h^k)}}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_h^k,V_{h+1})}+BS^2A\right) && \text{(Cauchy-Schwarz inequality)}\\
&=\tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}(P_h^k,V_{h+1})}+BS^2A\right).
\end{aligned}$$

Plugging these back completes the proof. □

Lemma 26. *(Shani et al., 2020, Lemma 1) For any policy $\pi\in(\Delta_{\mathcal{A}})^{\mathcal{S}\times[H]}$, two transition functions P,P', and utility function $d\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, we have $V_1^{\pi,P,d}(s)-V_1^{\pi,P',d}(s)=\mathbb{E}[\sum_{h=1}^{H}(P_{s_h,a_h,h}-P'_{s_h,a_h,h})V_{h+1}^{\pi,P,d}|\pi,P',s_1=s]$.*

Lemma 27. *(Chen et al., 2021a, Lemma 30) For a random variable X such that $|X|\le C$, we have: $\textsc{Var}[X^2]\le 4C^2\textsc{Var}[X]$.*

D. Concentration Inequalities

Lemma 28 (Any interval Azuma's inequality)**.** *Let $\{X_i\}_{i=1}^{\infty}$ be a martingale difference sequence and $|X_i|\le B$ almost surely. Then with probability at least $1-\delta$, for any l,n: $\left|\sum_{i=l}^{l+n-1}X_i\right|\le B\sqrt{2n\ln\frac{4(l+n-1)^3}{\delta}}$.*

Proof. For each $l,n\in\mathbb{N}_+$, we have with probability at least $1-\frac{\delta}{2(l+n-1)^3}$, $\left|\sum_{i=l}^{l+n-1}X_i\right|\le B\sqrt{2n\ln\frac{4(l+n-1)^3}{\delta}}$ by (Chen & Luo, 2021, Lemma 20). The statement is then proved by a union bound (note that $\sum_{l=1}^{\infty}\sum_{n=1}^{\infty}\frac{1}{2(l+n-1)^3}=\sum_{i=1}^{\infty}\sum_{j=1}^{i}\frac{1}{2i^3}=\sum_{i=1}^{\infty}\frac{1}{2i^2}\le 1$). □

Lemma 29. *(Chen et al., 2021b, Lemma 38) Let $\{X_i\}_{i=1}^{\infty}$ be a martingale difference sequence adapted to the filtration $\{\mathcal{F}_i\}_{i=0}^{\infty}$ and $|X_i|\le B$ for some $B>0$. Then with probability at least $1-\delta$, for all $n\ge 1$ simultaneously,*

$$\left|\sum_{i=1}^{n}X_i\right|\le 3\sqrt{\sum_{i=1}^{n}\mathbb{E}[X_i^2|\mathcal{F}_{i-1}]\ln\frac{4B^2n^3}{\delta}}+2B\ln\frac{4B^2n^3}{\delta}.$$

Lemma 30. *(Weissman et al., 2003) Given a distribution $p\in\Delta_m$ and let $\bar{p}$ be an empirical distribution of p over n samples. Then, $\|p-\bar{p}\|_1\le\sqrt{m\ln\frac{2}{\delta}/n}$ with probability at least $1-\delta$.*

Lemma 31. *(Cohen et al., 2020, Theorem D.3) Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d random variables with expectation μ and $X_n\in[0,B]$ almost surely. Then with probability at least $1-\delta$, for any $n\ge 1$:*

$$\left|\sum_{i=1}^{n}(X_i-\mu)\right|\le\min\left\{2\sqrt{B\mu n\ln\frac{2n}{\delta}}+B\ln\frac{2n}{\delta},2\sqrt{B\sum_{i=1}^{n}X_i\ln\frac{2n}{\delta}}+7B\ln\frac{2n}{\delta}\right\}.$$

Lemma 32. *(Cohen et al., 2020, Lemma D.4) and (Cohen et al., 2021, Lemma E.2) Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of random variables w.r.t to the filtration $\{\mathcal{F}_i\}_{i=0}^{\infty}$ and $X_i\in[0,B]$ almost surely. Then with probability at least $1-\delta$, for all $n\ge 1$*

simultaneously:

$$\sum_{i=1}^{n} \mathbb{E}[X_i|\mathcal{F}_{i-1}] \leq 2\sum_{i=1}^{n} X_i + 4B\ln\frac{4n}{\delta},$$
$$\sum_{i=1}^{n} X_i \leq 2\sum_{i=1}^{n} \mathbb{E}[X_i|\mathcal{F}_{i-1}] + 8B\ln\frac{4n}{\delta}.$$