LEARNING INFINITE-HORIZON AVERAGE-REWARD CMDPS VIA STATE AUGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study infinite-horizon average-reward constrained Markov decision processes (CMDPs) under the weakly communicating assumption. Existing high-probability guarantees for this setting either require computationally inefficient algorithms or have suboptimal dependence on the number of interactions T. We propose, to the best of our knowledge, the first computationally efficient algorithm that achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and cumulative constraint violation with high probability in the tabular setting. The $\sqrt{T}$ dependence is optimal up to logarithmic factors. Our approach incorporates cumulative constraint violation into the state and defines a reshaped reward through differences of a Huber potential. The added state determines the penalty on further violations while the reward function remains fixed on the augmented state space. Since the added state has known deterministic dynamics, only the original transition kernel needs to be estimated. The bounded slope of the Huber potential keeps the per-step reward bounded, and the potential differences telescope to relate the reshaped return to the original cumulative reward and the terminal potential. These properties allow us to apply finite-horizon approximation and optimistic value iteration with clipping, as used in unconstrained average-reward MDPs, without worsening the regret rate in T.

1 INTRODUCTION

In reinforcement learning (RL), actions that improve task performance may also increase resource consumption or the risk of undesirable outcomes (García & Fernández, 2015). For example, a robotic controller must account for safety costs while learning to perform its task (Achiam et al., 2017). Constrained Markov decision processes (CMDPs) model these requirements by maximizing expected reward subject to prescribed bounds on expected costs (Altman, 2021). When the transition dynamics are unknown, the learner must acquire information about the effects of its actions on both reward and constraint satisfaction. Its performance is therefore measured by regret relative to an optimal feasible policy and by the constraint violation accumulated during learning.

Learning in CMDPs has been studied extensively in the finite-horizon (Efroni et al., 2020; Qiu et al., 2020; Liu et al., 2021; Müller et al., 2024) and discounted settings (Chen et al., 2021b; Xu et al., 2021; Ding et al., 2023). The infinite-horizon average-reward setting remains less understood, although it directly describes continuing tasks whose performance is evaluated over long periods (Mahadevan, 1996; Sutton & Barto, 2018). In this setting, the agent interacts with the environment without resets, and both reward and constraint costs are evaluated through their long-run averages. For example, wireless transmission control can be formulated as minimizing average power consumption subject to a bound on average delay (Djonin & Krishnamurthy, 2007). The average-reward formulation captures these requirements without choosing a terminal horizon or a discount factor. This distinction also matters for feasibility: a bound on discounted cost need not enforce the corresponding bound on long-run average cost (Agnihotri et al., 2024).

We consider weakly communicating CMDPs, which allow stationary policies with multiple recurrent classes and are more general than ergodic and unichain CMDPs (Puterman, 1994). Learning in this setting inherits the difficulty of controlling bias spans in average-reward MDPs, since transition estimation errors depend on the variation of value functions across states. Algorithms for average-reward MDPs address this difficulty through span regularization or constraints (Bartlett & Tewari, 2009; Fruit et al., 2018), or through value function clipping (Hong et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Constraints on average costs, however, introduce further obstacles, making the analysis of weakly communicating CMDPs more challenging. An optimal constrained policy need not be greedy with respect to its reward action-value function, because a greedy policy may violate the constraint (Ghosh et al., 2023). Consequently, the Bellman optimality arguments used to control spans in unconstrained optimistic planning do not carry over directly (Chen et al., 2022). A common approach is to incorporate the constraint into the reward through a dual multiplier. However, updating this multiplier changes the reward used for planning and hence the value functions. Controlling this additional variation complicates the analysis of existing primal-dual algorithms (Ghosh et al., 2023; Yu et al., 2026). For example, in the finite-horizon approximation of Yu et al. (2026), updating the multiplier once per episode of length H yields a dual-regret bound of order $\sqrt{HT}$ after balancing the step-size terms. This term exceeds order $\sqrt{T}$ when H grows with T; see Appendix F.

In fact, existing results for weakly communicating CMDPs either require computationally inefficient algorithms or have suboptimal bounds on regret and constraint violation. In the tabular setting, Chen et al. (2022) solve a linear program under a finite-horizon approximation and achieves $\widetilde{\mathcal{O}}(T^{2/3})$ bounds on both quantities after T interactions. They give another algorithm that improves the bounds to $\widetilde{\mathcal{O}}(\sqrt{T})$ by adding span constraints, but these constraints make the planning problem nonconvex. For CMDPs with linear function approximation, Ghosh et al. (2023) obtain $\widetilde{\mathcal{O}}(\sqrt{T})$ bounds with a computationally inefficient algorithm, and $\widetilde{\mathcal{O}}(T^{3/4})$ bounds with an efficient primal-dual algorithm. More recently, Yu et al. (2026) establish strong duality for weakly communicating CMDPs and improve the efficient bounds in the linear setting to $\widetilde{\mathcal{O}}(T^{2/3})$.

Efficient $\widetilde{\mathcal{O}}(\sqrt{T})$ regret guarantees exist under stronger assumptions, including ergodicity (Agarwal et al., 2022a; Chen et al., 2022; Ghosh et al., 2023). Posterior-sampling algorithms also achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ Bayesian regret and constraint violation for ergodic or communicating CMDPs (Agarwal et al., 2022b; Provodin et al., 2024). Results for general policy parameterizations consider ergodic or unichain models and include approximation-error terms that can grow linearly with T (Bai et al., 2024; Satheesh & Aggarwal, 2026). These differences in assumptions and performance criteria are summarized in Table 1 in Appendix A.

These results leave open whether a computationally efficient algorithm can achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and cumulative constraint violation for weakly communicating average-reward CMDPs. In this work, we resolve this question in the tabular setting. Our contributions are summarized as follows.

- We propose State-Augmented Clipped Value Iteration with Upper Confidence Bound (SA-CVI-UCB, Algorithm 1) that achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and constraint violation with high probability (Theorem 2). To the best of our knowledge, this is the first computationally efficient algorithm with $\widetilde{\mathcal{O}}(\sqrt{T})$ guarantees for weakly communicating tabular average-reward CMDPs. The $\sqrt{T}$ dependence is optimal up to logarithmic factors when regret and constraint violation are controlled simultaneously (Jaksch et al., 2010; Singh et al., 2023). With a direct implementation of backward value iteration, the algorithm uses $\mathcal{O}(S^2AT^{5/2})$ arithmetic operations over T interactions.

- We develop a state-augmented formulation for regret analysis in average-reward CMDPs, combining a state variable that tracks cumulative constraint violation with a new reshaped reward based on a Huber potential. State augmentation has previously been used with safety budgets to enforce almost-sure constraints (Sootla et al., 2022b). In our formulation, the added state determines the penalty on further violations while the reward function on the augmented state space remains fixed. Since the transition of the added state is known and deterministic, only the original transition kernel needs to be estimated. We define the reshaped reward through differences of the Huber potential, whose bounded slope keeps the per-step reward bounded. The potential differences telescope, so the cumulative reshaped reward equals the original cumulative reward minus the terminal potential. Together, state augmentation and the reshaped reward allow us to apply techniques used for unconstrained average-reward MDPs without worsening the regret rate in T. Specifically, we combine finite-horizon approximation (Wei et al., 2021; Chen et al., 2022; Yu et al., 2026) with optimistic value iteration and value function clipping (Hong et al., 2025; Chae et al., 2025) to obtain $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and constraint violation in the original CMDP.

2 PROBLEM SETTING

Average-Reward CMDPs We consider a constrained Markov decision process (CMDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, s_1, r, g)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, respectively, P is the transition kernel, $s_1 \in \mathcal{S}$ is the initial state, and $r, g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are the reward and constraint functions, respectively. Here, $P(s' \mid s, a)$ denotes the probability of transitioning from state s to s' after taking action a. We assume that $r(s, a) \in [0, 1]$ and $g(s, a) \in [-1, 1]$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Moreover, we assume that r and g are known and deterministic, while the transition kernel P is unknown.

Let Π denote the set of stationary randomized policies. For any $\pi \in \Pi$ and initial state $s \in \mathcal{S}$, we define the average reward, or gain, as $J_r^\pi(s) = \lim_{T\to\infty}(1/T)\mathbb{E}_\pi[\sum_{t=1}^T r(s_t, a_t) \mid s_1 = s]$, where the expectation is taken over the trajectory generated by P and π. We define $J_g^\pi(s)$ analogously by replacing r with g. Given the initial state s_1, the average-reward CMDP is formulated as

$$\sup_{\pi\in\Pi} \quad J_r^\pi(s_1) \quad \text{s.t.} \quad J_g^\pi(s_1) \geq 0.$$

The agent interacts with the unknown CMDP for T steps, starting from s_1. At each step $t \in [T]$, the agent observes the current state s_t, chooses an action $a_t \in \mathcal{A}$ based on $a_t \sim \pi_t(\cdot|s_t)$, where π_t denotes the agent's policy. Subsequently, the agent observes the next state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. We evaluate the learning algorithm in terms of regret and constraint violation, defined as

$$\mathrm{Regret}(T) = \sum_{t=1}^T \big(J_r^* - r(s_t, a_t)\big), \qquad \mathrm{Violation}(T) = \sum_{t=1}^T \big(-g(s_t, a_t)\big).$$

Here, $\mathrm{Regret}(T)$ measures the cumulative reward gap relative to the optimal stationary policy, while $\mathrm{Violation}(T)$ measures the cumulative constraint violation. Finally, we impose the standard Slater condition, which ensures strict feasibility of the CMDP (Singh et al., 2023; Yu et al., 2026).

Assumption 1 (Slater condition)**.** There exists a stationary policy $\bar{\pi} \in \Pi$ such that $J_g^{\bar{\pi}}(s_1) \geq \gamma$ for some Slater constant $\gamma > 0$. We assume that γ is known, while the Slater policy $\bar{\pi}$ is unknown.

Weakly Communicating CMDPs Throughout the paper, we assume that the underlying MDP is weakly communicating. In particular, an MDP is weakly communicating if its state space can be partitioned into two sets such that all states in the first set communicate with one another under some stationary deterministic policies, while every state in the second set is transient under every stationary policy (Puterman, 1994). This assumption is weaker than the commonly imposed unichain and ergodic assumptions.

Unlike unconstrained weakly communicating MDPs, a weakly communicating CMDP does not necessarily admit an optimal stationary policy. Following previous works (Chen et al., 2022; Ghosh et al., 2023; Yu et al., 2026), we therefore impose the following assumption.

Assumption 2 (Existence of an optimal stationary policy)**.** There exists an optimal stationary policy $\pi^* \in \arg\max_{\pi\in\Pi}\{J_r^\pi(s_1) : J_g^\pi(s_1) \geq 0\}$ such that its reward and constraint gains are independent of the initial state. Namely, there exist constants J_r^* and J_g^* such that $J_r^{\pi^*}(s) = J_r^*$ and $J_g^{\pi^*}(s) = J_g^*$ for all $s \in \mathcal{S}$.

We denote by v_r^* and v_g^* the bias functions associated with π^* and the reward and constraint functions, respectively, defined as $v_r^*(s) = \lim_{T\to\infty}(1/T)\sum_{t=1}^T \mathbb{E}_{\pi^*}[\sum_{i=1}^t (r(s_i, a_i) - J_r^*) \mid s_1 = s]$ and $v_g^*(s) = \lim_{T\to\infty}(1/T)\sum_{t=1}^T \mathbb{E}_{\pi^*}[\sum_{i=1}^t (g(s_i, a_i) - J_g^*) \mid s_1 = s]$. For any $V : \mathcal{S} \to \mathbb{R}$, we define its span as $\mathrm{sp}(V) = \max_{s\in\mathcal{S}} V(s) - \min_{s\in\mathcal{S}} V(s)$.

3 ALGORITHM

In this section, we propose a computationally efficient algorithm for learning weakly communicating average-reward CMDPs that achieves rate-optimal guarantees in T. Our algorithm can be viewed as casting the original CMDP into an augmented MDP and then applying a learning algorithm to it. Importantly, the main technical novelty lies in the design of the augmented MDP: (i) state augmentation for regret analysis in the average-reward setting, and (ii) a new reshaped reward based on a Huber potential.

Before presenting our approach, we discuss the limitations of occupancy-measure optimization and primal-dual methods for weakly communicating average-reward CMDPs. Existing high-probability guarantees in this setting either require computationally inefficient planning or have suboptimal dependence on T. For instance, Algorithm 4 in Chen et al. (2022) achieves rate-optimal guarantees by imposing explicit constraints on the spans of the value functions induced by the occupancy measures. These constraints make the feasible region generally nonconvex, and no efficient method for solving the resulting optimization problem is known. This motivates an approach that is both computationally efficient and rate-optimal in T.

Another widely used approach is to incorporate the constraint into the reward through a dual multiplier z_k, leading to the composite reward $r + z_k g$—called the primal-dual method. However, its direct extension to the average-reward setting becomes nontrivial. In particular, rate-optimal algorithms for unconstrained average-reward MDPs rely critically on controlling the deviation of value functions across iterations (Hong & Tewari, 2025). Once the dual multiplier is introduced, the variation of z_k induces additional variation in the composite reward and hence in the value functions, which makes the same rate-optimal analysis difficult in the constrained setting; see Section F for more details. This suggests that incorporating the constraint information into the reward leads to suboptimal guarantees due to the additional variation in the composite reward.

3.1 State-Augmented MDP

Motivated by the failure of primal-dual approaches, our method can be summarized as follows: to avoid the additional variation induced by incorporating the constraint information into the reward, we instead incorporate it into the state space, inspired by Sootla et al. (2022b). Intuitively, this avoids the additional variation while preserving the desired statistical guarantees, since the evolution of the safety state is predictable from the current augmented state and action; see (1).

Technically, the augmented (unconstrained) MDP $\widetilde{\mathcal{M}} = (\widetilde{\mathcal{S}}, \mathcal{A}, \widetilde{P}, (s_1, z_1), \tilde{r})$ is defined as follows. We introduce a safety state z_t and augment the original state space to $\widetilde{\mathcal{S}} = \mathcal{S} \times \mathcal{Z}$, where $\mathcal{Z} = \{\Delta n : n \in \mathbb{Z}, |\Delta n| \le 2T\}$ with $\Delta = 1/\sqrt{T}$. Given $z_1 = 0$, the safety state evolves as $z_{t+1} = \Pi_{\mathcal{Z}}(z_t - g(s_t, a_t))$, where $\Pi_{\mathcal{Z}}$ returns the closest element in $\mathcal{Z}$. For simplicity, let $\psi(s, z, a) = \Pi_{\mathcal{Z}}(z - g(s, a))$. We then define the augmented transition kernel and reshaped reward as

$$\widetilde{P}(s', z' \mid s, z, a) = P(s' \mid s, a) \cdot \mathbb{1}\{z' = \psi(s, z, a)\}, \tag{1}$$

$$\tilde{r}(s, z, a) = r(s, a) + \Phi_W(z) - \Phi_W(\psi(s, z, a)), \tag{2}$$

for each $(s, z, a) \in \mathcal{S} \times \mathcal{Z} \times \mathcal{A}$ and $(s', z') \in \mathcal{S} \times \mathcal{Z}$. This defines an unconstrained MDP over the augmented state space $\widetilde{\mathcal{S}}$. Here, $\Phi_W : \mathbb{R} \to \mathbb{R}_+$ is a nondecreasing potential function, whose formal definition will be provided later.

The safety state z_t and the reshaped reward $\tilde{r}$ can be interpreted as follows. Intuitively, z_t tracks the cumulative constraint violation up to step t: ignoring the projection, $z_t \approx -\sum_{\tau=1}^{t-1} g(s_\tau, a_\tau)$, so a large positive z_t indicates accumulated constraint violation, whereas a negative z_t indicates accumulated constraint surplus. Accordingly, since Φ_W is nondecreasing, the potential difference term in (2) acts as a regularization term for constraint satisfaction: actions that decrease the safety state receive a nonnegative regularization term (i.e., a bonus), whereas actions that increase it receive a nonpositive one (i.e., a penalty). Thus, this term plays a role analogous to the penalty induced by a dual multiplier, with the magnitude of the regularization determined by the current safety state. Consequently, optimizing the reshaped reward naturally balances reward maximization and constraint satisfaction, motivating an unconstrained learning problem over the augmented MDP.

Motivation of $\tilde{r}$ We describe the design motivation for $\tilde{r}$, as it plays a critical role in our analysis. From a technical perspective, $\tilde{r}$ provides a natural bridge from the regret analysis of the augmented MDP to that of the original CMDP. In particular, its potential-difference form in (2) admits the telescoping structure, i.e., $\sum_{t=1}^{T} \big(\Phi_W(z_t) - \Phi_W(z_{t+1})\big) = -\Phi_W(z_{T+1})$, since $z_1 = 0$ and $\Phi_W(0) = 0$. Based on this, the cumulative reshaped reward can be expressed as

$$\sum_{t=1}^{T} \tilde{r}(s_t, z_t, a_t) = \sum_{t=1}^{T} r(s_t, a_t) - \Phi_W(z_{T+1}). \tag{3}$$

This relation allows us to translate a regret bound for the augmented MDP, expressed in terms of $\tilde{r}(s_t, z_t, a_t)$, into a bound on the original regret together with the terminal potential, which serves as a useful ingredient for controlling both regret and violation; see Step 2 of Section 4.

Choice of Φ_W: Huber Potential Previously, we discussed that the potential-difference structure in the design of $\tilde{r}$ provides a bridge between the augmented MDP and the original CMDP in the analysis. The remaining question is whether we can successfully carry out the regret analysis for the augmented MDP. To this end, several regularity conditions are typically required; for example, the reshaped reward $\tilde{r}$ should be sufficiently bounded. However, for common choices of potential functions such as quadratic or exponential potentials, $\tilde{r}$ may become large because these potentials are not globally Lipschitz. In particular, since the safety state z can grow with T, the potential difference $\Phi(z) - \Phi(\psi(s, z, a))$ may also become large, leading to an unfavorable dependence on T in the regret analysis.

This observation motivates us to choose a potential function with favorable regularity properties, such as Lipschitzness. For this purpose, we use the Huber potential Φ_W, defined as follows: for constants $\Lambda, W > 0$,

$$\Phi_W(x) = \begin{cases} 0, & x \le 0, \\ \Lambda x^2/(2W), & 0 < x \le W, \\ \Lambda\left(x - W/2\right), & x > W. \end{cases} \tag{4}$$

For comparison of potential functions, since $\Phi_W(x)$ grows linearly when $x > W$, it is globally Lipschitz, whereas quadratic and exponential potentials are not. Besides, the Huber potential satisfies additional properties that are useful in our analysis, which are summarized in the following lemma.

Lemma 1. *Let $\Phi_W : \mathbb{R} \to \mathbb{R}$ be defined as in (4). Then we have the following properties:*

1. $0 \le \Phi'_W(x) \le \Lambda$ *for all* $x \in \mathbb{R}$.
2. $\Phi'_W(y)x \le \Phi_W(x + y) - \Phi_W(y) \le \Phi'_W(y)x + \Lambda x^2/(2W)$ *for all* $x, y \in \mathbb{R}$.
3. *For any* $\lambda \in [0, \Lambda)$ *and* $x \in \mathbb{R}$, *we have* $(\Lambda - \lambda)[x]_+ \le \Phi_W(x) - \lambda x + \Lambda W/2$.

Comparison of Augmented RL While our state augmentation is related to that of Sootla et al. (2022b), the main difference in the resulting augmented MDP lies in the reshaped reward. In both approaches, the augmented state tracks cumulative constraint information, up to a difference in the sign convention for the safety state z_t. For the reshaped reward, Sootla et al. (2022b) imposes an infinite regularization once the constraint is violated. In a reward-maximization form, their reshaped objective can be expressed as

$$\tilde{r}(s, z, a) = \begin{cases} r(s, a), & z \ge 0, \\ -\infty, & z < 0, \end{cases}$$

thereby targeting almost-sure constraint satisfaction. In contrast, our reshaped reward (2) uses the potential difference $\Phi_W(z) - \Phi_W(\psi(s, z, a))$, which remains uniformly bounded due to the Lipschitzness of the Huber potential. This choice is tailored to our regret analysis, allowing us to control the augmented value functions while retaining the telescoping structure needed to bound regret and constraint violation.

3.2 Description of Algorithm

We present the proposed State-Augmented Clipped Value Iteration with Upper Confidence Bound (SA-CVI-UCB) in Algorithm 1. The algorithm adopts the finite-horizon approximation framework with K episodes and horizon H, so that $T = KH$. Accordingly, we use the trajectory notation s_h^k, a_h^k instead of s_t, a_t, with the correspondence $t = (k-1)H + h$. The algorithmic parameter choices, including K and H, are summarized in (5). In each episode $k \in [K]$, the algorithm first performs backward value iteration over the augmented state space $\widetilde{\mathcal{S}} = \mathcal{S} \times \mathcal{Z}$ using the empirical transition kernel $\hat{P}_k$. It then executes the greedy policy for H steps while updating the safety state according to ψ, and finally updates the empirical transition model.

Algorithm 1 State-Augmented Clipped Value Iteration with UCB (SA-CVI-UCB)

Input: the optimism parameter β; the number of steps T; the horizon length H and $K = T/H$; the Slater constant γ; the clipping parameter C and span $\mathrm{sp}(v_r^*), \mathrm{sp}(v_g^*)$; the parameter for Lyapunov function W; the discretization gap Δ;
Initialize: $z_1^1 \leftarrow 0$; $V_{k,H+1}(s,z) \leftarrow 0$; $\mathcal{Z} = \{\Delta n : n \in \mathbb{Z}, |\Delta n| \le 2T\}$;
$\hat{P}_1(s'|s,a) \leftarrow 1/S$ and $N_1(s,a,s'), N_1(s,a) \leftarrow 0 \quad \forall (s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$;
$\tilde{r}(s,a,z) \leftarrow r(s,a) + \Phi_W(z) - \Phi_W(\psi(s,z,a)) \quad \forall (s,a,z) \in \mathcal{S}\times\mathcal{A}\times\mathcal{Z}$;

1: **for** $k = 1, \dots, K$ **do**
2: **for** $h = H, \dots, 1$ **do**
3: **for** $(s,a,z) \in \mathcal{S}\times\mathcal{A}\times\mathcal{Z}$ **do**
4: $Q_{k,h}(s,a,z) \leftarrow \tilde{r}(s,a,z) + \hat{P}_k(\cdot|s,a)^\top V_{k,h+1}(\cdot, \psi(s,a,z)) + \beta/\sqrt{N_k(s,a) \vee 1}$
5: $\widetilde{V}_{k,h}(s,z) \leftarrow \max_{a\in\mathcal{A}} Q_{k,h}(s,a,z)$
6: $V_{k,h}(s,z) \leftarrow \widetilde{V}_{k,h}(s,z) \wedge (\min_{s'\in\mathcal{S}} \widetilde{V}_{k,h}(s',z) + 2C)$
7: **end for**
8: **end for**
9: $N_{k+1}(\cdot,\cdot,\cdot) \leftarrow N_k(\cdot,\cdot,\cdot)$ and $N_{k+1}(\cdot,\cdot) \leftarrow N_k(\cdot,\cdot)$
10: **for** $h = 1, \dots, H$ **do**
11: Take $a_h^k \in \arg\max_{a\in\mathcal{A}} Q_{k,h}(s_h^k, a, z_h^k)$
12: Observe $s_{h+1}^k \sim P(\cdot|s_h^k, a_h^k)$
13: $N_{k+1}(s_h^k, a_h^k, s_{h+1}^k) \leftarrow N_{k+1}(s_h^k, a_h^k, s_{h+1}^k) + 1$
14: $N_{k+1}(s_h^k, a_h^k) \leftarrow N_{k+1}(s_h^k, a_h^k) + 1$
15: $z_{h+1}^k \leftarrow \psi(s_h^k, z_h^k, a_h^k)$
16: **end for**
17: $s_1^{k+1} \leftarrow s_{H+1}^k$ and $z_1^{k+1} \leftarrow z_{H+1}^k$
18: $\hat{P}_{k+1}(s'|s,a) \leftarrow \begin{cases} N_{k+1}(s,a,s')/N_{k+1}(s,a) & \text{if } N_k(s,a) > 0, \\ 1/S & \text{if } N_k(s,a) = 0, \end{cases} \quad \forall (s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$
19: **end for**

In Lines 4–6, the backward recursion constructs an optimistic value estimate for the augmented MDP. Importantly, the true transition kernel $P(\cdot \mid s,a)$ is unknown, while the next safety state $\psi(s,a,z)$ is deterministic. Hence, $\hat{P}_k$ is estimated using all visits to (s,a), independently of the safety state, and the bonus $\beta/\sqrt{N_k(s,a) \vee 1}$ accounts only for uncertainty in P.

In Line 6, for each fixed $z \in \mathcal{Z}$, the algorithm clips the value-function estimate across the original state space. As shown in Section 4, the comparator value function in the augmented MDP has span at most $2C$ for every fixed safety state. Therefore, clipping at this level preserves optimism while allowing the statistical error to depend on the tighter span parameter C, whereas without clipping one would rely on the naïve span bound of H for $\widetilde{V}_{k,h}$. This step is crucial for obtaining the desired regret bound.

After planning, in Lines 10–18, the algorithm executes the action $a_h^k \in \arg\max_a Q_{k,h}(s_h^k, a, z_h^k)$ and observes the next state s_{h+1}^k. The safety state is then updated deterministically as $z_{h+1}^k = \psi(s_h^k, a_h^k, z_h^k)$. Both components of the augmented state are carried across episode boundaries, and the empirical transition kernel is updated using the cumulative transition counts.

3.3 Main Result

We present our main result. We choose the algorithmic parameters as follows:

$$\begin{aligned} &\delta \in (0, 1/3),\ \Delta = 1/\sqrt{T},\ \Lambda = 2/\gamma,\ W = K = H = \sqrt{T}, \\ &C = \mathrm{sp}(v_r^*) + \Lambda\,\mathrm{sp}(v_g^*) + \frac{\Lambda}{W}\left(\mathrm{sp}(v_g^*)^2 + 2H(\mathrm{sp}(v_g^*) + 2)^2\right) + \frac{H\Lambda\Delta}{2}, \\ &\beta = 2CS\sqrt{\log(2TS^2|\mathcal{A}|/\delta)/2}. \end{aligned} \tag{5}$$

Moreover, let $\mathrm{sp}(v^*_{\lambda^*})$ denote the span of the optimal bias function with respect to the composite reward $r + \lambda^* g$. where λ^* is the optimal dual multiplier; see Appendix B. Note that this quantity appears only in the analysis and it not required by the algorithm. Then, we have the following result.

Theorem 2. *For any $\delta \in (0, 1/3)$, with probability at least $1 - 3\delta$, we have*

$$\mathrm{Regret}(T) \le \widetilde{\mathcal{O}}\left(\left(\mathrm{sp}(v_r^*) + \frac{1}{\gamma}(\mathrm{sp}(v_g^*) + \mathrm{sp}(v_g^*)^2 + 1)\right) S^2 A\sqrt{T}\right),$$
$$\mathrm{Violation}(T) \le \widetilde{\mathcal{O}}\left((\mathrm{sp}(v_r^*) + \mathrm{sp}(v_g^*) + \mathrm{sp}(v_g^*)^2) S^2 A\sqrt{T} + \mathrm{sp}(v^*_{\lambda^*})\sqrt{T}\right).$$

The theorem shows that SA-CVI-UCB achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and constraint violation, up to problem-dependent factors involving S, A, γ, and the relevant bias spans. In particular, these results establish SA-CVI-UCB as the first computationally efficient algorithm to achieve rate-optimal guarantees in this setting. Here, the running time of SA-CVI-UCB is polynomial in S, A, and T. Under the parameter choice (5), we have $|\mathcal{Z}| \le 4T^{3/2} + 1$, and in each episode, Lines 4–6 compute $Q_{k,h}(s, a, z)$ for all $(h, s, a, z) \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{Z}$ with $\mathcal{O}(S)$ operations each. Hence, the total running time over K episodes is $\mathcal{O}(KHS^2A|\mathcal{Z}|) = \mathcal{O}(S^2AT^{5/2})$.

Why does the augmentation technique work? Some readers may wonder, in principle, why the state-augmentation technique leads to rate-optimal guarantees in the average-reward setting. To address this, we highlight two useful properties of the augmented formulation in our setting. First, $\tilde{r}$ is a fixed reward function on the augmented state space, while the constraint information evolves through the Markovian safety state. Hence, the additional reward variation and value function deviation induced by a changing dual multiplier is avoided. Second, since g is known and deterministic, the transition of the safety state is also known. Therefore, in the tabular setting, state augmentation introduces no additional unknown transition dynamics, and the learner only needs to estimate the original transition kernel $P(\cdot \mid s, a)$. These properties make the augmented formulation particularly suitable for obtaining rate-optimal guarantees in the average-reward setting.

4 Analysis

In this section, we present the analysis of Theorem 2. To establish bounds on $\mathrm{Regret}(T)$ and $\mathrm{Violation}(T)$, our proof proceeds in the following three-step framework. The detailed proofs are deferred to Appendix C.

$$\underbrace{\mathrm{Regret}_{\mathrm{aug}}(T)}_{\textbf{Step 1}} \xRightarrow[\textbf{Step 2}]{} \mathrm{Regret}(T) + \Phi_W(z_{T+1}) \xRightarrow[\textbf{Step 3}]{} \mathrm{Regret}(T), \mathrm{Violation}(T).$$

First, we establish a regret bound for the augmented MDP under the reshaped reward $\tilde{r}$ and transition kernel $\widetilde{P}$. We denote this regret by $\mathrm{Regret}_{\mathrm{aug}}(T)$ and formally define it as $\mathrm{Regret}_{\mathrm{aug}}(T) = \sum_{k=1}^K (V^{\pi^*}_{\tilde{r},1}(s_1^k, z_1^k) - \sum_{h=1}^H \tilde{r}(s_h^k, a_h^k, z_h^k))$, where $V^{\pi^*}_{\tilde{r},1}(s, z) = \mathbb{E}_{\widetilde{P},\pi^*}[\sum_{h=1}^H \tilde{r}(s_h, a_h, z_h) \mid s_1 = s, z_1 = z]$. Second, we transfer this bound to the original CMDP to obtain a bound on $\mathrm{Regret}(T) + \Phi_W(z_{T+1})$. Finally, we use this bound to separately control $\mathrm{Regret}(T)$ and $\mathrm{Violation}(T)$.

Before going through each step, we introduce the key lemma underlying our analysis. Roughly, this lemma ensures that the two quantities—$\mathrm{sp}(V^{\pi^*}_{\tilde{r},h}(\cdot, z))$ and $HJ_r^* - V^{\pi^*}_{\tilde{r},1}(s, z)$—are small, as $C = \widetilde{\mathcal{O}}(1)$ under the parameter choice (5). For comparison, this lemma can be viewed as an analogue of Lemma 2 in Wei et al. (2020), which considers the simpler setting of average-reward MDPs without state augmentation.

Lemma 3. *Let C be defined in (5). Then we have, for all $(h, s, z) \in [H] \times \mathcal{S} \times \mathcal{Z}$,*

1. $\mathrm{sp}(V^{\pi^*}_{\tilde{r},h}(\cdot, z)) \le 2C$,
2. $HJ_r^* - V^{\pi^*}_{\tilde{r},1}(s, z) \le C$.

Each property plays a crucial role in establishing Steps 1 and 2, respectively. In particular, the first statement ensures that the benchmark value function appearing in $\mathrm{Regret}_{\mathrm{aug}}(T)$ has a uniformly

bounded span, which is essential for controlling the regret in the augmented MDP. Moreover, the second statement provides a bridge between $\mathrm{Regret}_{\text{aug}}(T)$ and $\mathrm{Regret}(T)$, since $V_{\tilde{r},1}^{\pi^*}$ and J_r^* serve as the respective benchmarks for these two regret metrics. Given this lemma, we are now ready to proceed through each step.

Step 1: Analysis of the Augmented MDP To bound $\mathrm{Regret}_{\text{aug}}(T)$, we first introduce the following decomposition:

$$\begin{aligned}\mathrm{Regret}_{\text{aug}}(T) \le &\sum_{k=1}^{K}\left(V_{\tilde{r},1}^{\pi^*}(s_1^k, z_1^k) - V_{k,1}(s_1^k, z_1^k)\right)\\ &+\sum_{k=1}^{K}\sum_{h=1}^{H}\left(P(\cdot|s_h^k, a_h^k)^\top V_{k,h+1}(\cdot, z_{h+1}^k) - V_{k,h+1}(s_{h+1}^k, z_{h+1}^k)\right)\\ &+\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{2\beta}{\sqrt{N_k(s_h^k, a_h^k)\vee 1}}.\end{aligned}$$

Here, the second and third summations on the right-hand side can be bounded by $\widetilde{\mathcal{O}}(C\sqrt{T})$ and $\widetilde{\mathcal{O}}(\beta\sqrt{SAT})$, respectively, using standard arguments. The first summation is nonpositive by the following optimism lemma.

Lemma 4 (Optimism)**.** *Let $\delta \in (0,1)$. With probability at least $1-\delta$, simultaneously for all $(k,h,s,z) \in [K]\times[H]\times\mathcal{S}\times\mathcal{Z}$, we have $V_{\tilde{r},h}^{\pi^*}(s,z) \le V_{k,h}(s,z)$.*

The first statement of Lemma 3 plays a crucial role in establishing this lemma. In particular, a standard UCB analysis ensures that $V_{\tilde{r},h}^{\pi^*}(s,z) \le \widetilde{V}_{k,h}(s,z)$, where $\widetilde{V}_{k,h}(s,z)$ denotes the unclipped value function estimate. The first statement of Lemma 3 then ensures that this optimism property is preserved even after clipping. As a consequence, combining these bounds yields

$$\mathrm{Regret}_{\text{aug}}(T) \le \widetilde{\mathcal{O}}(\sqrt{T}). \tag{6}$$

Step 2: Transfer to the Original CMDP Next, based on the bound on $\mathrm{Regret}_{\text{aug}}(T)$, we establish a bound on $\mathrm{Regret}(T) + \Phi_W(z_{T+1})$. To this end, by the telescoping structure induced by the design of $\tilde{r}$, we have

$$\mathrm{Regret}_{\text{aug}}(T) = \mathrm{Regret}(T) + \Phi_W(z_{T+1}) + \sum_{k=1}^{K}\left(V_{\tilde{r},1}^{\pi^*}(s_1^k, z_1^k) - HJ_r^*\right).$$

Here, the summation $\sum_{k=1}^{K}\left(V_{\tilde{r},1}^{\pi^*}(s_1^k, z_1^k) - HJ_r^*\right)$ on the right-hand side represents the conversion error between the augmented MDP and the original CMDP. By the second statement of Lemma 3, this error term can be lower bounded by $-CK$. Since $K = \sqrt{T}$, combining this with the bound on $\mathrm{Regret}_{\text{aug}}(T) \le \widetilde{\mathcal{O}}(\sqrt{T})$ yields

$$\mathrm{Regret}(T) + \Phi_W(z_{T+1}) \le \widetilde{\mathcal{O}}(\sqrt{T}). \tag{7}$$

Step 3: Obtaining $\mathrm{Regret}(T)$ and $\mathrm{Violation}(T)$ To conclude the analysis, we derive separate bounds on regret and constraint violation using (7). In particular, since Φ_W is nonnegative, the bound on $\mathrm{Regret}(T)$ follows directly. We therefore focus on obtaining a bound on $\mathrm{Violation}(T)$.

This step, however, is nontrivial for the Huber potential. For comparison, in the case of a quadratic or exponential potential, it suffices to naïvely lower bound $\mathrm{Regret}(T) \ge -T$. Combining this with (7) then yields $z_{T+1} \le \widetilde{\mathcal{O}}(\sqrt{T})$, which concludes the analysis since z_{T+1} approximates $\mathrm{Violation}(T)$ up to the discretization error ΔT. In contrast, since the Huber potential $\Phi_W(x)$ grows linearly when $x \ge W$, the same argument no longer yields the desired bound.

To overcome this issue, our strategy is as follows. Instead of using the naïve lower bound $\mathrm{Regret}(T) \ge -T$, we exploit a sharper lower bound derived from strong duality; see Appendix B for details. In particular,

$$-\lambda^* z_{T+1} \lesssim \mathrm{Regret}(T),$$

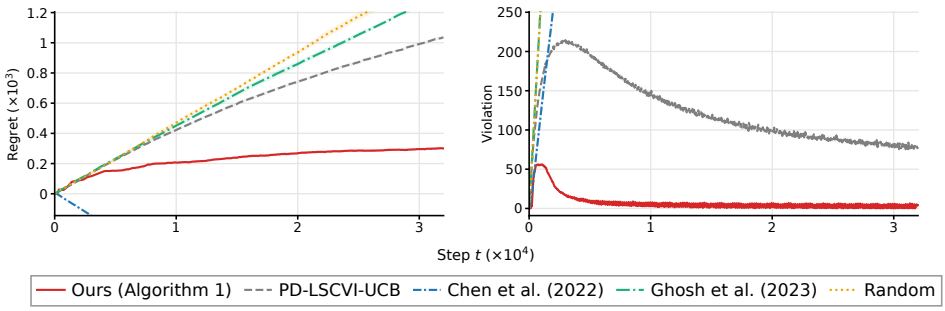


Figure 1: Plots of regret and constraint violation for $T = 32{,}000$ steps. Each plot represents the average over 10 trials, and shaded regions indicate 95% confidence intervals.

where λ^* denotes the optimal dual variable associated with the original CMDP, which is guaranteed to satisfy $\lambda^* \leq 1/\gamma$, and $\lesssim$ denotes an informal inequality used to emphasize intuition. Combining this lower bound with (7) yields

$$\Phi_W(z_{T+1}) - \lambda^* z_{T+1} \leq \widetilde{\mathcal{O}}(\sqrt{T}).$$

To derive a bound on z_{T+1} from this relation, we use the third statement of Lemma 1, which states that $(\Lambda - \lambda^*)[z_{T+1}]_+ \leq \Phi_W(z_{T+1}) - \lambda^* z_{T+1} + \Lambda W/2$. Combining this with the facts that $\Lambda - \lambda^* \geq 1/\gamma$ and $W = \sqrt{T}$ yields

$$z_{T+1} \leq \widetilde{\mathcal{O}}(\sqrt{T}). \tag{8}$$

Finally, again, since z_{T+1} approximates $\mathrm{Violation}(T)$ up to the discretization error, we obtain $\mathrm{Violation}(T) \leq \widetilde{\mathcal{O}}(\sqrt{T})$.

5 Numerical Experiments

We evaluate Algorithm 1 on a tabular CMDP obtained by modifying the CMDP of Yu et al. (2026). In the early states of the chain, the reward is given to the action that lowers the constraint value, so the policy that maximizes the reward violates the constraint and the algorithm must trade off the reward against the constraint. We compare Algorithm 1 with four baselines, (i) PD-LSCVI-UCB (Yu et al., 2026), (ii) Algorithm 3 of Chen et al. (2022), (iii) Algorithm 2 of Ghosh et al. (2023), and (iv) a random policy. We repeat 10 simulations with $T = 32{,}000$ steps and report the cumulative regret and constraint violation, which are summarized in Figure 1. Algorithm 1 achieves sublinear regret and constraint violation, whereas the constraint violation of Algorithm 2 of Ghosh et al. (2023) grows linearly. Compared with PD-LSCVI-UCB, Algorithm 1 achieves lower regret and constraint violation. Algorithm 3 of Chen et al. (2022) attains negative regret at the cost of a constraint violation that grows linearly.

We also examine the sensitivity of each algorithm to the scale of its bonus term, and find that Algorithm 1 achieves sublinear regret and constraint violation for every scale $c \leq 1$, whereas every baseline incurs linear regret or constraint violation for some c. The results are shown in Figure 2. Additional experiments and experimental details are deferred to Section E.

6 Conclusion

We studied infinite-horizon average-reward CMDPs under the weakly communicating assumption. Our algorithm, SA-CVI-UCB, is computationally efficient and achieves rate-optimal $\widetilde{\mathcal{O}}(T^{1/2})$ regret and constraint violation in the tabular setting. The main idea is to incorporate cumulative constraint violation into the state space and design the reshaped reward using a Huber potential. This formulation keeps the reshaped reward bounded while avoiding the additional reward variation caused by changing dual multipliers. The algorithm combines this formulation with finite-horizon approximation, optimistic value iteration, and value function clipping. It remains open whether the same computational efficiency and rate-optimal guarantees can be achieved for weakly communicating average-reward linear CMDPs. We refer the reader to Section F for a detailed discussion of this extension and the limitations of the primal-dual approach.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

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

Mridul Agarwal, Qinbo Bai, and Vaneet Aggarwal. Concave utility reinforcement learning with zero-constraint violations. *Transactions on Machine Learning Research*, 2022a.

Mridul Agarwal, Qinbo Bai, and Vaneet Aggarwal. Regret guarantees for model-based reinforcement learning with long-term average constraints. In *Conference on Uncertainty in Artificial Intelligence*, pp. 22–31. PMLR, 2022b.

Akhil Agnihotri, Rahul Jain, and Haipeng Luo. ACPO: A policy optimization algorithm for average MDPs with constraints. In *International Conference on Machine Learning*, pp. 397–415. PMLR, 2024.

Leopoldo Agorio, Sean Van Alen, Miguel Calvo-Fullana, Santiago Paternain, and Juan Andrés Bazerque. Multi-agent assignment via state augmented reinforcement learning. In *Learning for Dynamics and Control Conference*, volume 242, pp. 1202–1213. PMLR, 2024. URL https://proceedings.mlr.press/v242/agorio24a.html.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Santiago Amaya-Corredor, Miguel Calvo-Fullana, and Anders Jonsson. Scalable constrained multi-agent reinforcement learning via state augmentation and consensus for separable dynamics. *arXiv preprint arXiv:2605.30461*, 2026. URL https://arxiv.org/abs/2605.30461.

Qinbo Bai, Washim Uddin Mondal, and Vaneet Aggarwal. Learning general parameterized policies for infinite horizon average reward constrained MDPs via primal-dual policy gradient algorithm. In *Advances in Neural Information Processing Systems*, volume 37, pp. 108566–108599, 2024.

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Conference on Uncertainty in Artificial Intelligence*, pp. 35–42, 2009.

Victor Boone and Zihan Zhang. Achieving tractable minimax optimal regret in average reward MDPs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26728–26769, 2024.

Miguel Calvo-Fullana, Santiago Paternain, Luiz F. O. Chamon, and Alejandro Ribeiro. State augmented constrained reinforcement learning: Overcoming the limitations of learning with rewards. *IEEE Transactions on Automatic Control*, 69(7):4275–4290, 2024.

Nicolas Carrara, Edouard Leurent, Romain Laroche, Tanguy Urvoy, Odalric-Ambrym Maillard, and Olivier Pietquin. Budgeted reinforcement learning in continuous state space. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4fe5149039b52765bde64beb9f674940-Abstract.html.

Agustin Castellano, Hancheng Min, Juan Andrés Bazerque, and Enrique Mallada. Reinforcement learning with almost sure constraints. In *Proceedings of the 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pp. 559–570. PMLR, 2022. URL https://proceedings.mlr.press/v168/castellano22a.html.

Woojin Chae, Kihyuk Hong, Yufan Zhang, Ambuj Tewari, and Dabeen Lee. Learning infinite-horizon average-reward linear mixture MDPs of bounded span. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2737–2745. PMLR, 2025. URL https://proceedings.mlr.press/v258/chae25a.html.

Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward markov decision process with constraints. In *International Conference on Machine Learning*, pp. 3246–3270. PMLR, 2022.

Xiaoyu Chen, Jiachen Hu, Lihong Li, and Liwei Wang. Efficient reinforcement learning in factored MDPs with application to constrained RL. In *International Conference on Learning Representations*, 2021a. URL https://arxiv.org/abs/2008.13319.

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

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167): 1–51, 2018. URL https://jmlr.org/papers/v18/15-636.html.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-iterate convergent policy gradient primal-dual methods for constrained mdps. *Advances in Neural Information Processing Systems*, 36:66138–66200, 2023.

Dejan V. Djonin and Vikram Krishnamurthy. Q-learning algorithms for constrained Markov decision processes with randomized monotone policies: Application to MIMO transmission control. *IEEE Transactions on Signal Processing*, 55(5):2170–2181, 2007.

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

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pp. 1578–1586. PMLR, 2018.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

Arnob Ghosh and Mehrdad Moharrami. Online learning in risk sensitive constrained MDP. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19406–19425. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ghosh25c.html.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Achieving sub-linear regret in infinite horizon average reward constrained MDP with linear function approximation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zZhX4eYNeeh.

Edwin Hamel-De le Court, Francesco Belardinelli, and Alexander W. Goodall. Probabilistic shielding for safe reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16091–16099, 2025. doi: 10.1609/aaai.v39i15.33767. URL https://ojs.aaai.org/index.php/AAAI/article/view/33767.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 4259–4280. PMLR, 2022.

Kihyuk Hong and Ambuj Tewari. A computationally efficient algorithm for infinite-horizon average-reward linear MDPs. In *Forty-second International Conference on Machine Learning*, 2025.

Kihyuk Hong, Woojin Chae, Yufan Zhang, Dabeen Lee, and Ambuj Tewari. Reinforcement learning for infinite-horizon average-reward linear MDPs via approximation by discounted-reward MDPs. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=yNeoBG7WOP.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Hao Jiang, Tien Mai, Pradeep Varakantham, and Huy Hoang. Reward penalties on augmented states for solving richly constrained RL effectively. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19867–19875, 2024. doi: 10.1609/aaai.v38i18.29962. URL https://ojs.aaai.org/index.php/AAAI/article/view/29962.

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

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

Sridhar Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, 22(1–3):159–195, 1996.

Jeremy McMahan. Deterministic policies for constrained reinforcement learning in polynomial time. In *Advances in Neural Information Processing Systems*, volume 37, pp. 94453–94489, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ab9f9cfe97da3665e08f50ade9f8c4d6-Abstract-Conference.html.

Jeremy McMahan. Polynomial-time approximability of constrained reinforcement learning. In *International Conference on Machine Learning*, volume 267, pp. 43417–43439. PMLR, 2025a. URL https://proceedings.mlr.press/v267/mcmahan25b.html.

Jeremy McMahan. Anytime-constrained equilibria in polynomial time. In *International Conference on Machine Learning*, volume 267, pp. 43399–43416. PMLR, 2025b. URL https://proceedings.mlr.press/v267/mcmahan25a.html.

Jeremy McMahan and Xiaojin Zhu. Anytime-constrained reinforcement learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4321–4329. PMLR, 2024. URL https://proceedings.mlr.press/v238/mcmahan24a.html.

Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained mdps. In *International Conference on Machine Learning*, pp. 36605–36653. PMLR, 2024.

Danil Provodin, Maurits Clemens Kaptein, and Mykola Pechenizkiy. Efficient exploration in average-reward constrained reinforcement learning: Achieving near-optimal regret with posterior sampling. In *International Conference on Machine Learning*, pp. 41144–41162. PMLR, 2024.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. *Advances in Neural Information Processing Systems*, 33:15277–15287, 2020.

Anirudh Satheesh and Vaneet Aggarwal. Regret analysis of unichain average reward constrained MDPs with general parameterization. *arXiv preprint arXiv:2602.08000*, 2026.

Rahul Singh, Abhishek Gupta, and Ness B. Shroff. Learning in constrained Markov decision processes. *IEEE Transactions on Control of Network Systems*, 10(1):441–453, 2023. Preprint arXiv:2002.12435, 2020.

Aivar Sootla, Alexander Cowen-Rivers, Jun Wang, and Haitham Bou Ammar. Enhancing safe exploration using safety state augmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/debd0ae2083160397a22a4a8831c7230-Abstract-Conference.html.

Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David H Mguni, Jun Wang, and Haitham Ammar. Saute RL: Almost surely safe reinforcement learning using state augmentation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20423–20443. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/sootla22a.html.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, second edition, 2018.

Kaiwen Wang, Dawen Liang, Nathan Kallus, and Wen Sun. A reductions approach to risk-sensitive reinforcement learning with optimized certainty equivalents. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=kzoNjAtwby.

Zepeng Wang, Xiaochuan Shi, Chao Ma, Libing Wu, and Jia Wu. CCPO: Conservatively constrained policy optimization using state augmentation. In *ECAI 2023*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pp. 2599–2606. IOS Press, 2023. doi: 10.3233/FAIA230566. URL https://doi.org/10.3233/FAIA230566.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pp. 10170–10180. PMLR, 2020.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward MDPs with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3007–3015. PMLR, 2021.

Honghao Wei, Xin Liu, and Lei Ying. A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3868–3876, 2022.

Huan Xu and Shie Mannor. Probabilistic goal markov decision processes. In *International Joint Conference on Artificial Intelligence*, pp. 2046–2052, 2011. URL https://www.ijcai.org/Proceedings/11/Papers/341.pdf.

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

Yang Xu, Swetha Ganesh, Washim Uddin Mondal, Qinbo Bai, and Vaneet Aggarwal. Global convergence for average reward constrained MDPs with primal-dual actor critic algorithm. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/857af2ae549352c584b8f1953c38198d-Abstract-Conference.html.

Zhaoxing Yang, Haiming Jin, Yao Tang, and Guiyun Fan. Risk-aware constrained reinforcement learning with non-stationary policies. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 2029–2037. International Foundation for Autonomous Agents and Multiagent Systems, 2024. URL https://www.ifaamas.org/Proceedings/aamas2024/pdfs/p2029.pdf.

Kihyun Yu, Beomhan Baek, and Dabeen Lee. Learning weakly communicating average-reward cmdps: Strong duality and improved regret. *arXiv preprint arXiv:2605.11586*, 2026.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/9e984c108157cea74c894b5cf34efc44-Abstract.html.

Zihan Zhang and Qiaomin Xie. Sharper model-free reinforcement learning for average-reward Markov decision processes. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pp. 5476–5477. PMLR, 2023. URL https://proceedings.mlr.press/v195/zhang23b.html.

A Related Work

Table 1: Comparison of guarantees for learning infinite-horizon average-reward CMDPs with unknown transitions after T steps. Problem-dependent factors are suppressed. Bounds hold with high probability unless marked as expected or Bayesian guarantees. WC denotes weak communication; the WC results assume an optimal stationary policy with state-independent gains. Additional assumptions and differences in performance criteria are specified below.

Algorithm	Setting	Assumption	Regret	Violation	Efficient
Wei et al. (2022)[a]	Tabular	Ergodic	$\widetilde{\mathcal{O}}(T^{5/6})$	0	✓
Agarwal et al. (2022a)[a,b]	Tabular	Ergodic	$\widetilde{\mathcal{O}}(\sqrt{T})$	0	✓
Agarwal et al. (2022b)[c]	Tabular	Ergodic	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✓
Chen et al. (2022) (Alg. 1)	Tabular	Ergodic	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(1)$	✓
Singh et al. (2023)	Tabular	Unichain[d]	$\widetilde{\mathcal{O}}(T^{2/3})$	$\widetilde{\mathcal{O}}(T^{2/3})$	✓[e]
Ghosh et al. (2023) (Alg. 3)[f]	Linear	Ergodic	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✓
Provodin et al. (2024)[c,g]	Tabular	Communicating	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✓
Bai et al. (2024)[h]	General	Ergodic	$\widetilde{\mathcal{O}}(T^{4/5})$	$\widetilde{\mathcal{O}}(T^{4/5})$	✓
Satheesh & Aggarwal (2026)[i]	General	Unichain	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✓
Chen et al. (2022) (Alg. 3)	Tabular	WC	$\widetilde{\mathcal{O}}(T^{2/3})$	$\widetilde{\mathcal{O}}(T^{2/3})$	✓
Chen et al. (2022) (Alg. 4)	Tabular	WC	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✗
Ghosh et al. (2023) (Alg. 1)[j]	Linear	WC	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✗
Ghosh et al. (2023) (Alg. 2)	Linear	WC	$\widetilde{\mathcal{O}}(T^{3/4})$	$\widetilde{\mathcal{O}}(T^{3/4})$	✓
Yu et al. (2026)	Linear	WC	$\widetilde{\mathcal{O}}(T^{2/3})$	$\widetilde{\mathcal{O}}(T^{2/3})$	✓
Ours (Algorithm 1)	Tabular	WC	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✓

[a]Zero aggregate violation holds under the stated sufficiently-large-T conditions. For Wei et al. (2022), both bounds hold in expectation, and zero violation means nonpositive expected signed cumulative violation, not safety at every step. [b]The row reports UC-CURL, which covers Lipschitz concave objectives and convex constraints, including standard CMDPs, under a Slater condition. [c]Bayesian guarantees, averaging over the transition prior and interaction randomness, under the stated strict-feasibility and horizon conditions. [d]Bounds hold in expectation. In addition to unichain structure, the analysis uses uniform geometric convergence and a finite worst-policy state-to-state hitting time. [e]The optimistic planning problem admits an LP reformulation using state-action-next-state occupancy measures. [f]Requires uniform mixing and a uniformly positive definite stationary feature Gram matrix under every stationary policy (Assumptions 4 and 5). [g]Requires a prior supported on models with bounded diameter, a strictly feasible policy inducing an irreducible and aperiodic chain, and sufficiently large T. [h]Expected trajectory guarantees under general policy parameterization, with smoothness and Fisher non-degeneracy assumptions. The displayed rates omit the policy approximation term. [i]Expected guarantees under general policy parameterization, with bounded Lipschitz scores, Fisher non-degeneracy, critic approximation and uniform hitting-time assumptions. The displayed rates omit the policy and critic approximation terms. The violation bound uses average costs of policy iterates rather than costs realized along the trajectory. [j]Requires the optimal policy to belong to a class of soft-max policies.

Average-Reward CMDPs Average-reward CMDPs model continuing tasks in which an agent maximizes long-term average reward subject to long-term average constraints (Altman, 2021). Model-based methods establish finite-time regret guarantees for average-reward CMDPs (Singh et al., 2023; Chen et al., 2022). Model-free algorithms also achieve sublinear expected regret and nonpositive expected cumulative constraint violation for sufficiently large T (Wei et al., 2022). Under the ergodic assumption, Chen et al. (2022) obtain $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and $\widetilde{\mathcal{O}}(1)$ constraint violation. Under the weakly communicating assumption, their efficient finite-horizon approximation gives $\widetilde{\mathcal{O}}(T^{2/3})$ bounds for both quantities. They also achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ bounds by imposing span constraints on the finite-horizon value functions induced by the occupancy measure, but the resulting feasible region is generally nonconvex and no efficient solution method is known. For linear CMDPs, the efficient finite-horizon algorithm of Ghosh et al. (2023) obtains $\widetilde{\mathcal{O}}(T^{3/4})$ bounds. Yu et al. (2026) establish strong duality for weakly communicating CMDPs and improve these bounds to $\widetilde{\mathcal{O}}(T^{2/3})$ through primal-dual clipped value iteration. Posterior-sampling methods also provide Bayesian regret guarantees for communicating CMDPs (Provodin et al., 2024). Under the ergodic

assumption, general policy parameterizations have been studied through primal-dual policy gradient (Bai et al., 2024) and actor-critic methods (Xu et al., 2025). Satheesh & Aggarwal (2026) analyze actor-critic methods under the unichain assumption. Our work obtains computational efficiency together with $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and constraint violation with high probability for weakly communicating tabular CMDPs.

Our algorithm also builds on techniques for unconstrained average-reward MDPs. In the tabular setting, optimistic planning gives regret guarantees for communicating and weakly communicating MDPs (Jaksch et al., 2010; Bartlett & Tewari, 2009; Zhang & Ji, 2019; Boone & Zhang, 2024). Fruit et al. (2018) obtain efficient exploration under the weakly communicating assumption by controlling the bias span. Model-free approaches combine optimistic Q-learning with discounted-reward approximation (Wei et al., 2020; Zhang & Xie, 2023). For linear MDPs, Wei et al. (2021) develop efficient value-iteration algorithms based on finite-horizon approximation. Hong et al. (2025) combine discounted-reward approximation with value function clipping to achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. Hong & Tewari (2025) further remove the dependence of the computational complexity on the state-space size through efficient clipping and deviation-controlled value iteration. We adapt optimistic value iteration and value function clipping to a state-augmented MDP. The augmented formulation incorporates cumulative constraint violation into the state and avoids the additional reward variation caused by changing dual multipliers, which complicates the primal-dual extension of these techniques to CMDPs. More broadly, unconstrained average-reward MDPs have been studied extensively, with computationally efficient algorithms achieving minimax-optimal regret up to logarithmic factors in the tabular setting. We refer the reader to Boone & Zhang (2024) and the references therein for further discussion of this literature.

CMDPs via State Augmentation State augmentation allows policies to use constraint information that is not contained in the original state. One approach incorporates Lagrange multipliers into the state space. Calvo-Fullana et al. (2024) develop this approach for average-reward CMDPs, learning policies conditioned on the multipliers and updating the multipliers during execution to obtain asymptotic feasibility and near-optimality in expectation. This formulation retains dual updates and has also been extended to distributed multi-agent problems (Agorio et al., 2024; Amaya-Corredor et al., 2026). Another approach augments the state with a budget or cumulative cost. For expected discounted-cost constraints, Carrara et al. (2019) augment the state and action with a budget and establish a budgeted Bellman optimality equation. For episodic RL with knapsack constraints and finitely supported costs on a discrete grid, Chen et al. (2021a) track the remaining budgets and exploit the resulting factored transition structure to obtain $\widetilde{\mathcal{O}}(\sqrt{T})$ regret with near-optimal dependence on S, A, H, and T. The computational cost depends exponentially on the number of constraints.

Budget augmentation has also been studied under several notions of safety. Sauté RL (Sootla et al., 2022b) incorporates a remaining safety budget into the state and reshapes the objective to target almost-sure constraint satisfaction. Under the stated regularity assumptions, its optimal value functions converge to those of the infinite-penalty formulation as the violation penalty tends to infinity. Simmer (Sootla et al., 2022a) studies how to schedule the safety budget during training under expected-cost and almost-sure constraints, with empirical improvements in safe exploration. For terminating MDPs with binary damage, Castellano et al. (2022) characterize the minimum safety budget. They give sample-complexity guarantees for learning this budget from a generative model, assuming a positive lower bound on nonzero probabilities in the joint transition and damage kernel. For anytime constraints, which must hold at every step almost surely, McMahan & Zhu (2024) use cumulative-cost augmentation to derive planning and learning algorithms, together with hardness and approximation results. Related constructions support equilibrium computation in anytime-constrained Markov games (McMahan, 2025b). For probabilistic state-avoidance constraints, Hamel-De le Court et al. (2025) construct a shield on an augmented MDP that guarantees safety and preserves the constrained optimum when the safety dynamics are known. Using value-demand augmentation, McMahan (2024) compute feasible policies that are near-optimal among deterministic policies under a time-space-recursive constraint. McMahan (2025a) use artificial budget variables to obtain bicriteria approximation guarantees that allow a small constraint violation. These works study different feasibility criteria or planning objectives from the regret and cumulative violation considered here.

Risk Criteria via State Augmentation State augmentation also helps handle nonlinear criteria applied to cumulative rewards or costs. Xu & Mannor (2011) track accumulated reward in probabilistic-goal and chance-constrained MDPs. For conditional value-at-risk constraints, Chow et al. (2018) develop actor-critic methods on an augmented MDP and prove almost-sure convergence to locally optimal policies under their sampling and approximation assumptions. Yang et al. (2024) instead condition policies on the quantile level of accumulated cost. For episodic CMDPs with an entropic-risk constraint, Ghosh & Moharrami (2025) introduce a continuous budget variable and use a primal-dual algorithm to obtain $\widetilde{\mathcal{O}}(K^{3/4})$ regret and constraint violation over K episodes with high probability. Wang et al. (2025) reduce finite-horizon learning with optimized certainty equivalent objectives to risk-neutral learning in an augmented MDP.

Reward Penalties on Augmented States Closely related to our reward design, Wang et al. (2023) combine safety-state augmentation with state-dependent reward penalties and an adaptive penalty multiplier in the discounted setting. Jiang et al. (2024) combine cumulative-cost augmentation with trajectory-dependent reward penalties in finite-horizon problems. They give penalty conditions under which optimal policies of the augmented MDP satisfy expected-cost or chance constraints, and relate a modified penalty to excess-loss constraints. The discounted reward adjustments in Jiang et al. (2024) also admit a telescoping representation. We study continuing, weakly communicating average-reward CMDPs and track cumulative constraint violation throughout learning. Our reshaped reward uses a Huber potential whose bounded derivative keeps the per-step reward adjustment bounded. We use this property and a quadratic remainder bound to control the augmented finite-horizon value functions under an optimal policy of the original CMDP. In particular, we bound their span across the original states uniformly over the added state and control the shortfall from the original average-reward benchmark (Lemma 3). These bounds allow us to apply finite-horizon approximation and optimistic value iteration with clipping. Together with a strong-duality argument, this yields a computationally efficient algorithm with $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and constraint violation with high probability.

B Strong Duality

In this section, we introduce several lemmas that are associated with the strong duality of weakly communicating average-reward CMDPs (Yu et al., 2026), which will be used for Step 3 of Section 4. We have the following results.

Lemma 5 (Theorem 1 in Yu et al. (2026)). *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, g)$ be a weakly communicating average-reward CMDP with $S, |\mathcal{A}| < \infty$, and let $s_1 \in \mathcal{S}$ be any initial state. Suppose that the CMDP $\mathcal{M}$ is feasible. Then the following properties hold:*

1. *There exists $\lambda^* \geq 0$ such that $D(\lambda^*) = \inf_{\lambda \geq 0} D(\lambda)$.*
2. *Strong duality, i.e., $J_r^*(s_1) = D(\lambda^*)$.*

By utilizing the strong duality result, we can show the following lemma, which states optimal dual variable is bounded in terms of the Slater constant γ.

Lemma 6 (Lemma 3 in Yu et al. (2026)). *Suppose that Slater's condition holds, i.e., there exists a stationary policy $\bar{\pi}$ such that $J_g^{\bar{\pi}}(s_1) \geq \gamma$ for some $\gamma > 0$. Let λ^* denote the optimal dual variable defined in Lemma 5. Then $\lambda^* \leq 1/\gamma$.*

Given a bounded optimal dual variable λ^*, which is guaranteed by Lemmas 5 and 6, we consider an unconstrained average-reward MDP with respect to the reward function $r + \lambda^* g$. Then, the Bellman optimality condition implies the following: there exists $J_{\lambda^*}^* \in \mathbb{R}$ and $v_{\lambda^*}^* : \mathcal{S} \to \mathbb{R}$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$J_{\lambda^*}^* + v_{\lambda^*}^*(s) \geq r(s, a) + \lambda^* g(s, a) + P v_{\lambda^*}^*(s, a) \tag{9}$$

Finally, we introduce the following lemma that states when Slater's condition holds, then $\mathrm{Regret}(T)$ can be lower bounded in terms of $\mathrm{Violation}(T)$.

Lemma 7 (Lemma 7 in Yu et al. (2026)). *Let $v_{\lambda^*}^*$ be defined in (9). With probability at least $1 - \delta$,*

$$\mathrm{Regret}(T) \geq -\lambda^* \,\mathrm{Violation}(T) - \mathrm{sp}(v_{\lambda^*}^*)\sqrt{2T \log(1/\delta)} - \mathrm{sp}(v_{\lambda^*}^*).$$

C Deferred Proofs of Sections 3 and 4

C.1 Proof of Lemma 1

Lemma 8 (Restatement of Lemma 1)**.** *Let $\Phi_W : \mathbb{R} \to \mathbb{R}$ be defined as in (4). Then we have the following properties:*

1. *$0 \le \Phi'_W(x) \le \Lambda$ for all $x \in \mathbb{R}$.*
2. *$\Phi_W(x+y) - \Phi_W(y) \le \Phi'_W(y)x + \frac{\Lambda}{2W}x^2$ for all $x, y \in \mathbb{R}$.*
3. *$\Phi'_W(y)x \le \Phi_W(x+y) - \Phi_W(y)$ for all $x, y \in \mathbb{R}$.*
4. *For any $\lambda \in [0, \Lambda)$ and $x \in \mathbb{R}$, we have $\Phi_W(x) - \lambda x \ge (\Lambda - \lambda)[x]_+ - \frac{\Lambda W}{2}$.*

Proof. The first statement is clear, since we have

$$\Phi'_W(x) = \begin{cases} 0 & \text{if } x \le 0, \\ (\Lambda/W)x & \text{if } 0 < x \le W, \\ \Lambda & \text{if } W < x. \end{cases} \tag{10}$$

We prove the second and third statements. By (10), we have for any $x, y \in \mathbb{R}$ such that $x \le y$,

$$0 \le \Phi'_W(y) - \Phi'_W(x) \le \frac{\Lambda}{W}(y - x) \tag{11}$$

While $0 \le \Phi'_W(y) - \Phi'_W(x)$ is trivial, as Φ'_W is nondecreasing, the $\Phi'_W(y) - \Phi'_W(x) \le (\Lambda/W)(y - x)$ can be shown by the following arguments. When $x, y \le 0$ or $W \le x, y$, the statement is trivial. When $x \le 0 < y \le W$, we have $\Phi'_W(y) - \Phi'_W(x) = (\Lambda/W)y \le (\Lambda/W)(y - x)$. When $x < 0 \le W < y$, we have $\Phi'_W(y) - \Phi'_W(x) = \Lambda \le (\Lambda/W)(y - x)$. Finally, when $0 < x \le W < y$, we have $\Phi'_W(y) - \Phi'_W(x) \le \Lambda - (\Lambda/W)x \le (\Lambda/W)(y - x)$. Therefore, (11) is true. Then it follows that

$$0 \le \int_0^1 \left(\Phi'_W(y + \tau x) - \Phi'_W(y)\right) x d\tau \le \int_0^1 \frac{\Lambda}{W} \tau x^2 d\tau = \frac{\Lambda}{2W} x^2$$

where the inequality follows from (10). Moreover, $\int_0^1 \left(\Phi'_W(y + \tau x) - \Phi'_W(y)\right) x d\tau$ can be rewritten as follows.

$$\begin{aligned} \int_0^1 \left(\Phi'_W(y + \tau x) - \Phi'_W(y)\right) x d\tau &= \int_0^1 \Phi'_W(y + \tau x) x d\tau - \Phi'_W(y)x \\ &= \int_0^x \Phi'_W(y + \tau') d\tau' - \Phi'_W(y)x \\ &= \Phi_W(y + x) - \Phi_W(y) - \Phi'_W(y)x \end{aligned}$$

where the second equality follows from the substitution $\tau' = x\tau$, and the last equality follows from the fundamental theorem of calculus. Combining these leads to

$$0 \le \Phi_W(y + x) - \Phi_W(y) - \Phi'_W(y)x \le \frac{\Lambda}{2W} x^2.$$

Note that this implies the second and third statements.

Consider $\lambda \in [0, \Lambda)$. To prove the fourth statement, we consider the following three cases: (i) $x \le 0$, (ii) $0 < x \le W$, and (iii) $W < x$. For (i), we have $\Phi_W(x) - \lambda x = -\lambda x \ge 0 \ge -\frac{\Lambda W}{2} = (\Lambda - \lambda)[x]_+ - \frac{\Lambda W}{2}$. For (ii), we have

$$\Phi_W(x) - \lambda x - (\Lambda - \lambda)[x]_+ + \frac{\Lambda W}{2} = \frac{\Lambda}{2W} x^2 - \Lambda x + \frac{\Lambda W}{2} = \frac{\Lambda}{2W}(x - W)^2 \ge 0.$$

Therefore, $\Phi_W(x) - \lambda x \ge (\Lambda - \lambda)[x]_+ - \frac{\Lambda W}{2}$ when (ii). For (iii), we have

$$\Phi_W(x) - \lambda x = \Lambda(x - \frac{W}{2}) - \lambda x = (\Lambda - \lambda)[x]_+ - \frac{\Lambda W}{2}.$$

Finally, these imply that the third statement holds. □

C.2 Proof of Lemma 3

We first prove the following lemma, which is useful to show Lemma 3.

Lemma 9. *Define C as*

$$C = \mathrm{sp}(v_r^*) + \Lambda\,\mathrm{sp}(v_g^*) + \frac{\Lambda}{W}\left(\mathrm{sp}(v_g^*)^2 + 2H(\mathrm{sp}(v_g^*)+2)^2\right) + \frac{H\Lambda\Delta}{2}. \tag{12}$$

Then we have, for all $(h,s,z) \in [H]\times\mathcal{S}\times\mathcal{Z}$,

$$\left|(H-h+1)J_r^* + \Phi_W(z) - \Phi_W\left(z-(H-h+1)J_g^*\right) - V_{\tilde{r},h}^{\pi^*}(s,z)\right| \le C.$$

Proof. Recall that $\tilde{P}(s',z'|s,a,z) = \mathbb{1}\{z' = \Pi_{\mathcal{Z}}(z-g(s,a))\}\cdot P(s'|s,a)$. Hence, given $s_{h'}, a_{h'}, z_{h'}$, we have $z_{h'+1} = \Pi_{\mathcal{Z}}(z_{h'} - g(s_{h'},a_{h'}))$ for each h'. Then it follows that

$$\begin{aligned}
&V_{\tilde{r},h}^{\pi^*}(s,z)\\
&= \mathbb{E}_{\tilde{P},\pi^*}\left[\sum_{h'=h}^{H}\tilde{r}(s_{h'},a_{h'},z_{h'})|s_h=s,z_h=z\right]\\
&= \mathbb{E}_{\tilde{P},\pi^*}\left[\sum_{h'=h}^{H}\left(r(s_{h'},a_{h'}) + \Phi_W(z_{h'}) - \Phi_W\left(\Pi_{\mathcal{Z}}(z_{h'}-g(s_{h'},a_{h'}))\right)\right)|s_h=s,z_h=z\right]\\
&= \mathbb{E}_{\tilde{P},\pi^*}\left[\sum_{h'=h}^{H}\left(r(s_{h'},a_{h'}) + \Phi_W(z_{h'}) - \Phi_W(z_{h'+1})\right)|s_h=s,z_h=z\right]\\
&= \mathbb{E}_{\tilde{P},\pi^*}\left[\sum_{h'=h}^{H} r(s_{h'},a_{h'}) + \Phi_W(z_h) - \Phi_W(z_{H+1})|s_h=s,z_h=z\right].
\end{aligned} \tag{13}$$

Moreover, since $z_{h'+1} = \Pi_{\mathcal{Z}}(z_{h'}-g(s_{h'},a_{h'}))$, we have for all $h'\in[H]$,

$$|z_{h'+1} - (z_{h'} - g(s_{h'},a_{h'}))| \le \frac{\Delta}{2}.$$

Then it follows that

$$\begin{aligned}
\left|z_{H+1} - \left(z_h - \sum_{h'=h}^{H} g(s_{h'},a_{h'})\right)\right| &= \left|\sum_{h'=h}^{H}\left(z_{h'+1} - (z_{h'} - g(s_{h'},a_{h'}))\right)\right|\\
&\le \sum_{h'=h}^{H}|z_{h'+1} - (z_{h'} - g(s_{h'},a_{h'}))|\\
&\le \frac{(H-h+1)\Delta}{2}.
\end{aligned}$$

It leads to

$$\begin{aligned}
\left|\Phi_W(z_{H+1}) - \Phi_W\left(z_h - \sum_{h'=h}^{H} g(s_{h'},a_{h'})\right)\right| &\le \Lambda\left|z_{H+1} - \left(z_h - \sum_{h'=h}^{H} g(s_{h'},a_{h'})\right)\right|\\
&\le \frac{(H-h+1)\Lambda\Delta}{2}
\end{aligned}$$

where the first inequality follows from the Lipschitzness of Φ_W, i.e., $|\Phi_W(y)-\Phi_W(x)| \le \Lambda|y-x|$. Then we have

$$\begin{aligned}
\mathbb{E}_{\tilde{P},\pi^*}\left[\Phi_W(z_{H+1})|s_h=s,z_h=z\right] \le \mathbb{E}_{\tilde{P},\pi^*}&\left[\Phi_W\left(z_h - \sum_{h'=h}^{H} g(s_{h'},a_{h'})\right)|s_h=s,z_h=z\right]\\
&+ \frac{(H-h+1)\Lambda\Delta}{2}.
\end{aligned} \tag{14}$$

By applying (14) to (13), we have

$$\begin{aligned} V^{\pi^*}_{\tilde{r},h}(s,z) &\leq \mathbb{E}_{\tilde{P},\pi^*}\left[\sum_{h'=h}^{H} r(s_{h'},a_{h'})|s_h=s, z_h=z\right] \\ &\quad + \mathbb{E}_{\tilde{P},\pi^*}\left[\Phi_W(z_h) - \Phi_W\left(z_h - \sum_{h'=h}^{H} g(s_{h'},a_{h'})\right)|s_h=s, z_h=z\right] \\ &\quad + \frac{(H-h+1)\Lambda\Delta}{2} \\ &= \mathbb{E}_{P,\pi^*}\left[\sum_{h'=h}^{H} r(s_{h'},a_{h'})|s_h=s\right] \\ &\quad + \Phi_W(z) - \mathbb{E}_{P,\pi^*}\left[\Phi_W\left(z - \sum_{h'=h}^{H} g(s_{h'},a_{h'})\right)|s_h=s\right] \\ &\quad + \frac{(H-h+1)\Lambda\Delta}{2} \end{aligned} \tag{15}$$

where the second equality is true since terms in the expectations do not include $\{z_{h'}\}_{h'=h+1}^{H+1}$, it can be rewritten using P rather than $\tilde{P}$ along with the fact that $\mathbb{E}_{\tilde{P},\pi^*}[\Phi_W(z_h)|z_h=z] = \Phi_W(z)$. By applying (15) to the desired term, we have

$$\begin{aligned} &\left|(H-h+1)J^*_r + \Phi_W(z) - \Phi_W(z-(H-h+1)J^*_g) - V^{\pi^*}_{\tilde{r},h}(s,z)\right| \\ &\leq \underbrace{\left|(H-h+1)J^*_r - \mathbb{E}_{P,\pi^*}\left[\sum_{h'=h}^{H} r(s_{h'},a_{h'})|s_h=s\right]\right|}_{\text{(I)}} \\ &\quad + \underbrace{\left|\mathbb{E}_{P,\pi^*}\left[\Phi_W\left(z - \sum_{h'=h}^{H} g(s_{h'},a_{h'})\right)|s_h=s\right] - \Phi_W(z-(H-h+1)J^*_g)\right|}_{\text{(II)}} \\ &\quad + \frac{(H-h+1)\Lambda\Delta}{2}. \end{aligned} \tag{16}$$

For (I), by Lemma 20, we have

$$\text{(I)} \leq \mathrm{sp}(v^*_r).$$

Next, we bound (II). To show this, recall that for all $s \in \mathcal{S}$,

$$J^*_g + v^*_g(s) = \sum_{a\in\mathcal{A}} \pi^*(a|s)(g(s,a) + Pv^*_g(s,a)).$$

Then it follows that for each $h' \in [H]$,

$$\mathbb{E}_{P,\pi^*}\left[\underbrace{g(s_{h'},a_{h'}) + v^*_g(s_{h'+1}) - v^*_g(s_{h'}) - J^*_g}_{\triangleq \xi_{h'}}|\mathcal{F}_{h'}\right] = 0 \tag{17}$$

where $\mathcal{F}_{h'} = \sigma(s_1,a_1,\ldots,a_{h'-1},s_{h'})$. Moreover, we define $\xi_{h'} = g(s_{h'},a_{h'}) + v^*_g(s_{h'+1}) - v^*_g(s_{h'}) - J^*_g$. By summing $\xi_{h'}$ over $h' = h,\ldots,H$, we have

$$\sum_{h'=h}^{H} \xi_{h'} = \sum_{h'=h}^{H} g(s_{h'},a_{h'}) + v^*_g(s_{H+1}) - v^*_g(s_h) - (H-h+1)J^*_g.$$

It can be rewritten as for any $z \in \mathcal{Z}$ and $h \in [H]$,

$$z - \sum_{h'=h}^{H} g(s_{h'},a_{h'}) + v^*_g(s_h) - v^*_g(s_{H+1}) + \sum_{h'=h}^{H} \xi_{h'} = z - (H-h+1)J^*_g.$$

Now, we apply the second statement of Lemma 8. Then it follows that

$$\begin{aligned}
&\Phi_W\left(z-\sum_{h'=h}^{H} g(s_{h'},a_{h'})\right)-\Phi_W(z-(H-h+1)J_g^*)\\
&\le \Phi_W'\left(z-(H-h+1)J_g^*\right)\left(v_g^*(s_{H+1})-v_g^*(s_h)-\sum_{h'=h}^{H}\xi_{h'}\right)\\
&\quad+\frac{\Lambda}{2W}\left(v_g^*(s_{H+1})-v_g^*(s_h)-\sum_{h'=h}^{H}\xi_{h'}\right)^2
\end{aligned}$$

By taking $\mathbb{E}_{P,\pi^*}[\cdot|s_h=s]$, we have

$$\begin{aligned}
&\mathbb{E}_{P,\pi^*}\left[\Phi_W\left(z-\sum_{h'=h}^{H} g(s_{h'},a_{h'})\right)|s_h=s\right]-\Phi_W(z-(H-h+1)J_g^*)\\
&\le \Phi_W'\left(z-(H-h+1)J_g^*\right)\underbrace{\mathbb{E}_{P,\pi^*}\left[v_g^*(s_{H+1})-v_g^*(s)-\sum_{h'=h}^{H}\xi_{h'}|s_h=s\right]}_{\text{(III)}}\\
&\quad+\frac{\Lambda}{2W}\underbrace{\mathbb{E}_{P,\pi^*}\left[\left(v_g^*(s_{H+1})-v_g^*(s_h)-\sum_{h'=h}^{H}\xi_{h'}\right)^2|s_h=s\right]}_{\text{(IV)}}.
\end{aligned}\tag{18}$$

Next, we bound (III) and (IV) individually. For (III), note that

$$\mathbb{E}_{P,\pi^*}\left[\sum_{h'=h}^{H}\xi_{h'}|s_h=s\right]=\mathbb{E}_{P,\pi^*}\left[\sum_{h'=h}^{H}\mathbb{E}_{P,\pi^*}[\xi_{h'}|\mathcal{F}_{h'}]|s_h=s\right]=0$$

where the first equality follows from the tower rule, and the second equality follows from (17). Moreover, we have $v_g^*(s_{H+1})-v_g^*(s)\le \mathrm{sp}(v_g^*)$ almost surely. Then we have

$$|\text{(III)}|\le \mathrm{sp}(v_g^*).$$

To bound (IV), we have

$$\begin{aligned}
\text{(IV)}&\le 2\mathbb{E}_{P,\pi^*}\left[\mathrm{sp}(v_g^*)^2+\left(\sum_{h'=h}^{H}\xi_{h'}\right)^2|s_h=s\right]\\
&=2\,\mathrm{sp}(v_g^*)^2+2\mathbb{E}_{P,\pi^*}\left[\sum_{h'=h}^{H}\xi_{h'}^2|s_h=s\right]\\
&\le 2\,\mathrm{sp}(v_g^*)^2+2(H-h+1)(\mathrm{sp}(v_g^*)+2)^2
\end{aligned}$$

where the first inequality follows from the fact that $(x+y)^2\le 2x^2+2y^2$ for all $x,y\in\mathbb{R}$, and the last inequality is due to $|\xi_{h'}|\le 2+\mathrm{sp}(v_g)^*$ for all $h'\in[H]$, as $J_g^*\in[-1,1]$ and $g(s_{h'},a_{h'})\in[-1,1]$. Note that the equality can be shown by the following argument.

$$\begin{aligned}
\mathbb{E}_{P,\pi^*}\left[\left(\sum_{h'=h}^{H}\xi_{h'}\right)^2|s_h=s\right]&=\mathbb{E}_{P,\pi^*}\left[\sum_{h'=h}^{H}\xi_{h'}^2+\sum_{\substack{h\le h'',h'''\le H\\ h''\neq h'''}}\xi_{h''}\xi_{h'''}|s_h=s\right]\\
&=\mathbb{E}_{P,\pi^*}\left[\sum_{h'=h}^{H}\xi_{h'}^2|s_h=s\right]
\end{aligned}$$

where the second equality follows from for any h'' and h''' such that $h'' < h'''$, we have $\mathbb{E}_{P,\pi^*}[\xi_{h''}\xi_{h'''}] = \mathbb{E}_{P,\pi^*}[\xi_{h''}\mathbb{E}_{P,\pi^*}[\xi_{h'''}|\mathcal{F}_{h'''}]] = 0$ by the tower rule. By applying bounds on (III) and (IV) to (18), we have

$$\begin{aligned}
&\mathbb{E}_{P,\pi^*}\left[\Phi_W\left(z - \sum_{h'=h}^{H} g(s_{h'}, a_{h'})\right)|s_h = s\right] - \Phi_W(z - (H-h+1)J_g^*) \\
&\leq \Phi_W'\left(z - (H-h+1)J_g^*\right)|(\text{III})| + (\text{IV}) \\
&\leq \Lambda\,\mathrm{sp}(v_g^*) + \frac{\Lambda}{W}\left(\mathrm{sp}(v_g^*)^2 + 2(H-h+1)(\mathrm{sp}(v_g^*)+2)^2\right).
\end{aligned} \tag{19}$$

By applying the same argument with the third statement of Lemma 8, we have

$$\mathbb{E}_{P,\pi^*}\left[\Phi_W\left(z - \sum_{h'=h}^{H} g(s_{h'}, a_{h'})\right)|s_h = s\right] - \Phi_W(z - (H-h+1)J_g^*) \geq -\Lambda\,\mathrm{sp}(v_g^*) \tag{20}$$

where the only differences are the direction of inequality, and the absence of (IV). Then by (19) and (20), we have

$$\begin{aligned}
(\text{II}) &= \left|\mathbb{E}_{P,\pi^*}\left[\Phi_W\left(z - \sum_{h'=h}^{H} g(s_{h'}, a_{h'})\right)\right] - \Phi_W(z - (H-h+1)J_g^*)\right| \\
&\leq \Lambda\,\mathrm{sp}(v_g^*) + \frac{\Lambda}{W}\left(\mathrm{sp}(v_g^*)^2 + 2(H-h+1)(\mathrm{sp}(v_g^*)+2)^2\right).
\end{aligned}$$

Finally, by applying bounds on (I) and (II) to (16), we have

$$\begin{aligned}
&\left|(H-h+1)J_r^* + \Phi_W(z) - \Phi_W(z - (H-h+1)J_g^*) - V_{\tilde{r},h}^{\pi^*}(s,z)\right| \\
&\leq \mathrm{sp}(v_r^*) + \Lambda\,\mathrm{sp}(v_g^*) + \frac{\Lambda}{W}\left(\mathrm{sp}(v_g^*)^2 + 2(H-h+1)(\mathrm{sp}(v_g^*)+2)^2\right) + \frac{(H-h+1)\Lambda\Delta}{2} \\
&\leq C.
\end{aligned}$$

This concludes the proof of the first statement. □

Lemma 10 (Restatement of Lemma 3)**.** *Let C defined in (12). Then we have, for all $(h,s,z) \in [H]\times\mathcal{S}\times\mathcal{Z}$,*

1. $\mathrm{sp}(V_{\tilde{r},h}^{\pi^*}(\cdot,z)) \leq 2C$,
2. $HJ_r^* - V_{\tilde{r},1}^{\pi^*}(s,z) \leq C$.

Proof. We prove the first statement. For a fixed $z \in \mathcal{Z}$ and any $s', s'' \in \mathcal{S}$, we have

$$\begin{aligned}
&|V_{\tilde{r},h}^{\pi^*}(s',z) - V_{\tilde{r},h}^{\pi^*}(s'',z)| \\
&\leq \left|(H-h+1)J_r^* + \Phi_W(z) - \Phi_W(z-(H-h+1)J_g^*) - V_{\tilde{r},h}^{\pi^*}(s',z)\right| \\
&\quad + \left|(H-h+1)J_r^* + \Phi_W(z) - \Phi_W(z-(H-h+1)J_g^*) - V_{\tilde{r},h}^{\pi^*}(s'',z)\right| \\
&\leq 2C
\end{aligned}$$

where the first inequality follows from the triangle inequality, and the second inequality follows from (9). Then it follows that

$$\mathrm{sp}(V_{\tilde{r},h}^{\pi^*}(\cdot,z)) = \max_{s',s''}\left|V_{\tilde{r},h}^{\pi^*}(s',z) - V_{\tilde{r},h}^{\pi^*}(s'',z)\right| \leq 2C.$$

Next, we prove the second statement. By Lemma 9,

$$(H-h+1)J_r^* + \Phi_W(z) - \Phi_W(z-(H-h+1)J_g^*) - V_{\tilde{r},h}^{\pi^*}(s,z) \leq C.$$

Note that since $J_g^* \geq 0$, we have $z \geq z - (H-h+1)J_g^*$. Then since Φ_W is a nondecreasing function, we have $\Phi_W(z) - \Phi_W(z-(H-h+1)J_g^*) \geq 0$. Therefore, it follows that

$$(H-h+1)J_r^* - V_{\tilde{r},h}^{\pi^*}(s,z) \leq C.$$

□ □

C.3 Deferred proofs for Step 1 of Section 4

Lemma 11. *Suppose that $\Delta \le 1$. We have $z_h^k \in \mathcal{Z}$ for all $(k,h) \in [K] \times [H]$.*

Proof. For all $(k,h) \in [K] \times [H]$, recall that $z_{h+1}^k = \Pi_{\mathcal{Z}}(z_h^k - g(s_h^k, a_h^k))$. Moreover, note that

$$\begin{aligned}
|z_{h+1}^k - z_h^k| &= \left|\Pi_{\mathcal{Z}}(z_h^k - g(s_h^k, a_h^k)) - z_h^k\right| \\
&= \left|\Pi_{\mathcal{Z}}(z_h^k - g(s_h^k, a_h^k)) - (z_h^k - g(s_h^k, a_h^k)) + g(s_h^k, a_h^k)\right| \\
&\le \left|\Pi_{\mathcal{Z}}(z_h^k - g(s_h^k, a_h^k)) - (z_h^k - g(s_h^k, a_h^k))\right| + \left|g(s_h^k, a_h^k)\right| \\
&\le \frac{\Delta}{2} + 1
\end{aligned}$$

where the last inequality follows from the error due to the projection onto the grid with gap Δ is at most $\Delta/2$, and the fact that $g(s_h^k, a_h^k) \in [-1, 1]$. Recall that $z_{H+1}^k = z_h^{k+1}$ and $z_1^1 = 0$. Then we have for all $(k,h) \in [K] \times [H]$,

$$\begin{aligned}
|z_{h+1}^k| &= \left|\sum_{k=1}^{K}\sum_{h=1}^{H}(z_{h+1}^k - z_h^k)\right| \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H}\left|z_{h+1}^k - z_h^k\right| \\
&\le \frac{\Delta T}{2} + T \\
&\le 2T.
\end{aligned}$$

Then it implies that $|z_h^k| \le 2T$ for all $(k,h) \in [K] \times [H]$. Since z_h^k is obtained from $\Pi_{\mathcal{Z}}$, we have

$$z_h^k \in \{\Delta n : n \in \mathbb{Z}, |\Delta n| \le 2T\} = \mathcal{Z}.$$

□

Lemma 12. *Suppose that $S \ge 2$. With probability at least $1-\delta$, for all $(k,h,s,a,z) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{Z}$, we have*

$$\left|\left(\hat{P}_k(\cdot|s,a) - P(\cdot|s,a)\right)^\top V_{k,h+1}(\cdot, \psi(s,a,z))\right| \le \frac{\beta}{\sqrt{N_k(s,a) \vee 1}}$$

where $\beta = 2CS\sqrt{\log(2TS^2|\mathcal{A}|/\delta)/2}$.

Proof. Fix $(k,h,s,a,z) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{Z}$. By the algorithm, we have $\sum_{s'} \hat{P}_k(s'|s,a) = 1$. Then we have

$$\left|\left(\hat{P}_k(\cdot|s,a) - P(\cdot|s,a)\right)^\top V_{k,h+1}(\cdot, \psi(s,a,z))\right| \le 2C\left\|\hat{P}_k(\cdot|s,a) - P(\cdot|s,a)\right\|_1 \tag{21}$$

where the inequality follows from Lemma 19, and the fact that $\mathrm{sp}(V_{k,h}(\cdot, \psi(s,a,z))) \le 2C$.

Next, we consider the following two cases: (i) $N_k(s,a) = 0$ and (ii) $N_k(s,a) > 0$. For (i), the desired statement becomes trivial, i.e., when $S \ge 2$,

$$2C\left\|\hat{P}_k(\cdot|s,a) - P(\cdot|s,a)\right\|_1 \le 4C \le \frac{\beta}{\sqrt{N_k(s,a) \vee 1}}.$$

For (ii), it can be shown as follows. Fix k, s, a, s'. Let $n = N_k(s,a)$, and define the random process $\{X_i\}_{i=1}^n$ such that

$$X_i = \begin{cases} 1 & \text{if } s' \text{ is visited immediately after the } i\text{th visit to } (s,a), \\ 0 & \text{otherwise.} \end{cases}$$

Then, $\{X_i\}_{i=1}^n$ are i.i.d. random variables that satisfy $\mathbb{E}[X_i] = P(s'|s,a)$ for all i. By Hoeffding's inequality, with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^n X_i - P(s'|s,a)\right| \le \sqrt{\frac{\log(2/\delta)}{2n}}.$$

By letting $\delta \leftarrow \delta/(TS^2|\mathcal{A}|)$ and applying a union bound over n, s, a, s', we have

$$\left|\frac{1}{n}\sum_{i=1}^n X_i - P(s'|s,a)\right| \le \sqrt{\frac{\log(2TS^2|\mathcal{A}|/\delta)}{2n}}, \ \ \forall (n,s,a,s') \in [T]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}.$$

Note that the above inequality holds for all $n \in [T]$, we can substitute $n = N_k(s,a)$. Moreover, we have $(1/n)\sum_{i=1}^n X_i = N_k(s,a,s')/(N_k(s,a)\vee 1) = \hat{P}_k(s'|s,a)$. Thus, it follows that

$$|\hat{P}_k(s'|s,a) - P(s'|s,a)| \le \sqrt{\frac{\log(2TS^2|\mathcal{A}|/\delta)}{2N_k(s,a)}}, \ \ \forall (k,s,a,s') \in [K]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}.$$

By summing this over $s' \in \mathcal{S}$, we have

$$\left\|\hat{P}_k(\cdot|s,a) - P(\cdot|s,a)\right\|_1 \le S\sqrt{\frac{\log(2KS^2|\mathcal{A}|/\delta)}{2N_k(s,a)}}, \ \ \forall (k,s,a) \in [K]\times\mathcal{S}\times\mathcal{A}.$$

Finally, for both cases (i) and (ii), we have shown that, with probability at least $1-\delta$,

$$2C\left\|\hat{P}_k(\cdot|s,a) - P(\cdot|s,a)\right\|_1 \le 2CS\sqrt{\frac{\log(2KS^2|\mathcal{A}|/\delta)}{2(N_k(s,a)\vee 1)}}, \ \ \forall (k,s,a) \in [K]\times\mathcal{S}\times\mathcal{A}$$

By applying this to (21), with probability at least $1-\delta$, for $(k,s,a) \in [K]\times\mathcal{S}\times\mathcal{A}$, we have

$$\left|\left(\hat{P}_k(\cdot|s,a) - P(\cdot|s,a)\right)^\top V_{k,h+1}(\cdot,\psi(s,a,z))\right| \le 2CS\sqrt{\frac{\log(2KS^2|\mathcal{A}|/\delta)}{2(N_k(s,a)\vee 1)}}$$

as desired. □

Lemma 13 (Restatement of Lemma 4)**.** *Suppose that Lemma 12 holds. For any $(k,h,s,z) \in [K]\times[H]\times\mathcal{S}\times\mathcal{Z}$, we have*

$$V^{\pi^*}_{\tilde{r},h}(s,z) \le V_{k,h}(s,z).$$

Proof. Fix $k \in [K]$. We prove the statement by induction on $h = H+1,\dots,1$. For the base case, since $V_{k,H+1}(s,z)$ is initialized to be 0, we have $V^{\pi^*}_{\tilde{r},H+1}(s,z) = V_{k,H+1}(s,z) = 0$ for all $(s,z) \in \mathcal{S}\times\mathcal{Z}$. Now, we assume that $V^{\pi^*}_{\tilde{r},h+1}(s,z) \le V_{k,h+1}(s,z)$ for all $(s,z) \in \mathcal{S}\times\mathcal{Z}$. For simplicity, define the auxiliary function ψ such that

$$\psi(s,a,z) = \Pi_{\mathcal{Z}}(z - g(s,a)).$$

Note that

$$\begin{aligned} Q^{\pi^*}_{\tilde{r},h}(s,a,z) &= \tilde{r}(s,a,z) + \sum_{s'\in\mathcal{S},z'\in\mathbb{R}} \tilde{P}(s',z'|s,a,z)V^{\pi^*}_{\tilde{r},h+1}(s',z') \\ &= \tilde{r}(s,a,z) + \sum_{s'\in\mathcal{S}} P(s'|s,a)V^{\pi^*}_{\tilde{r},h+1}(s',\psi(s,a,z)) \end{aligned}$$

where the first equality follows from the definition of the Q-function, and the last equality follows from the definition of $\tilde{P}(s',z'|s,a,z) = \mathbb{1}\{z' = \Pi_{\mathcal{Z}}(z-g(s,a))\}P(s'|s,a) = \mathbb{1}\{z' = \psi(s,a,z)\}P(s'|s,a)$. Then it follows that

$$\begin{aligned} &Q_{k,h}(s,a,z) - Q^{\pi^*}_{\tilde{r},h}(s,a,z) \\ &= \hat{P}_k(\cdot|s,a)^\top V_{k,h+1}(\cdot,\psi(s,a,z)) - P(\cdot|s,a)^\top V^{\pi^*}_{\tilde{r},h+1}(s',\psi(s,a,z)) + \frac{\beta}{\sqrt{N_k(s,a)}} \\ &= \left(\hat{P}_k(\cdot|s,a) - P(\cdot|s,a)\right)^\top V_{k,h+1}(\cdot,\psi(s,a,z)) \\ &\quad + P(\cdot|s,a)^\top\left(V_{k,h+1}(\cdot,\psi(s,a,z)) - V^{\pi^*}_{\tilde{r},h+1}(\cdot,\psi(s,a,z))\right) + \frac{\beta}{\sqrt{N_k(s,a)}} \\ &\ge 0 \end{aligned}$$

where the inequality follows from Lemma 12, i.e., $(\hat{P}_k(\cdot|s,a) - P(\cdot|s,a))^\top V_{k,h+1}(\cdot, \psi(s,a,z)) \geq -\beta/\sqrt{N_k(s,a)}$, and the induction hypothesis, i.e., $V_{k,h+1}(s,z) \geq V^{\pi^*}_{\tilde{r},h+1}(s,z)$. Then we have for all $(s,z) \in \mathcal{S} \times \mathcal{Z}$,

$$\widetilde{V}_{k,h}(s,z) = \max_{a \in \mathcal{A}} Q_{k,h}(s,a,z) \geq \sum_{a \in \mathcal{A}} \pi^*(a|s) Q^{\pi^*}_{\tilde{r},h}(s,a,z) = V^{\pi^*}_{\tilde{r},h}(s,z). \tag{22}$$

Recall that $V_{k,h}(s,z) = \widetilde{V}_{k,h}(s,z) \wedge (\min_{s'} \widetilde{V}_{k,h}(s',z) + 2C)$ for all $z \in \mathcal{Z}$ by the algorithm. Then it follows that for all $(s,z) \in \mathcal{S} \times \mathcal{Z}$,

$$\begin{aligned} V_{k,h}(s,z) &= \widetilde{V}_{k,h}(s,z) \wedge (\min_{s'} \widetilde{V}_{k,h}(s',z) + 2C) \\ &\geq V^{\pi^*}_{\tilde{r},h}(s,z) \wedge (\min_{s'} V^{\pi^*}_{\tilde{r},h}(s',z) + 2C) \\ &= V^{\pi^*}_{\tilde{r},h}(s,z) \end{aligned}$$

where the inequality follows from (22), and the last equality follows from the first statement of Lemma 10. This concludes the induction, implying that $V_{k,h}(s,z) \geq V^{\pi^*}_{\tilde{r},h}(s,z)$ for all $(h,s,z) \in [H] \times \mathcal{S} \times \mathcal{Z}$. Since we can apply the same argument for all k, we conclude the proof. □

Lemma 14. *With probability at least $1-\delta$, we have*

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \left(P(\cdot|s_h^k, a_h^k)^\top V_{k,h+1}(\cdot, z_{h+1}^k) - V_{k,h+1}(s_{h+1}^k, z_{h+1}^k) \right) \leq 2C\sqrt{2T \log(2/\delta)}.$$

Proof. Recall that $\mathcal{F}_{k,h} = \sigma(s_1, a_1, \ldots, s_h^k, a_h^k)$. Note that z_{h+1}^k is $\mathcal{F}_{k,h}$-measurable, and $V_{k,h+1}$ is $\mathcal{F}_{k-1,H}$-measurable. Then it follows that

$$P(\cdot|s_h^k, a_h^k)^\top V_{k,h+1}(\cdot, z_{h+1}^k) = \mathbb{E}\left[V_{k,h+1}(s_{h+1}^k, z_{h+1}^k) | \mathcal{F}_{k,h} \right].$$

Define $X_{k,h} = P(\cdot|s_h^k, a_h^k)^\top V_{k,h+1}(\cdot, z_{h+1}^k) - V_{k,h+1}(s_{h+1}^k, z_{h+1}^k)$, which satisfies $\mathbb{E}[X_{k,h}|\mathcal{F}_{k,h}] = 0$. Thus, $\{X_{k,h}\}_{k \in [K], h \in [H]}$ is a martingale difference sequence. Then by the Azuma-Hoeffding inequality, with probability at least $1-\delta$, we have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} X_{k,h} \leq 2C\sqrt{2T \log(2/\delta)}.$$

□

Lemma 15. *We have*

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{2\beta}{\sqrt{N_k(s_h^k, a_h^k) \vee 1}} \leq 4\beta HSA + 6\beta\sqrt{SAT}.$$

Proof. Note that

$$\begin{aligned} \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{\sqrt{N_k(s_h^k, a_h^k) \vee 1}} &= \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{\sqrt{N_k(s_h^k, a_h^k) \vee 1}} \\ &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{k=1}^{K} \frac{\sum_{h=1}^{H} \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{N_k(s,a) \vee 1}}. \end{aligned}$$

Note that

$$\begin{aligned} &\sum_{k=1}^{K} \frac{\sum_{h=1}^{H} \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{N_k(s,a) \vee 1}} \\ &= \sum_{k: N_k(s,a) < H} \frac{\sum_{h=1}^{H} \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{N_k(s,a) \vee 1}} + \sum_{k: N_k(s,a) \geq H} \frac{\sum_{h=1}^{H} \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{N_k(s,a) \vee 1}} \\ &= \sum_{k: N_k(s,a) < H} \frac{\sum_{h=1}^{H} \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{N_k(s,a) \vee 1}} + \sum_{k: N_k(s,a) \geq H} \frac{\sum_{h=1}^{H} \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{N_k(s,a) \vee 1}}. \end{aligned} \tag{23}$$

We bound the first term. For simplicity, let $k_0 = \max\{k \in [K] : N_k(s,a) < H\}$. Recall that $N_k(s,a)$ denotes the number of visits to (s,a) up to $k-1$ episode. Then, we know that

$$\begin{aligned}\sum_{k:N_k(s,a)<H} \frac{\sum_{h=1}^H \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{N_k(s,a) \vee 1}} &\leq \sum_{k:N_k(s,a)<H} \sum_{h=1}^H \mathbb{1}\{s_h^k = s, a_h^k = a\} \\ &= \sum_{k=1}^{k_0} \sum_{h=1}^H \mathbb{1}\{s_h^k = s, a_h^k = a\} \\ &= \sum_{k=1}^{k_0-1} \sum_{h=1}^H \mathbb{1}\{s_h^k = s, a_h^k = a\} + \sum_{h=1}^H \mathbb{1}\{s_h^{k_0} = s, a_h^{k_0} = a\} \\ &\leq N_{k_0}(s,a) + H \\ &\leq 2H\end{aligned} \tag{24}$$

where the first inequality follows from $\sqrt{N_k(s,a) \vee 1} \geq 1$, the first equality follows from the definition of k_0, the second inequality follows from the definition of $N_{k_0}(s,a)$ and the fact that $\sum_{h=1}^H \mathbb{1}\{s_h^{k_0} = s, a_h^{k_0} = a\} \leq H$, and the last inequality is again due to the definition of k_0.

We then bound the second term. For each k such that $N_k(s,a) \geq H$, we have

$$\begin{aligned}N_{k+1}(s,a) &= N_k(s,a) + \sum_{h=1}^H \mathbb{1}\{s_h^k = s, a_h^k = a\} \\ &\leq N_k(s,a) + H \\ &\leq 2N_k(s,a).\end{aligned} \tag{25}$$

Then it follows that

$$\begin{aligned}\sum_{k:N_k(s,a)\geq H} \frac{\sum_{h=1}^H \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{N_k(s,a) \vee 1}} &= \sum_{k:N_k(s,a)\geq H} \frac{\sum_{h=1}^H \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{N_k(s,a)}} \\ &= 3 \sum_{k:N_k(s,a)\geq H} \frac{\sum_{h=1}^H \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{4N_k(s,a)} + \sqrt{N_k(s,a)}} \\ &\leq 3 \sum_{k:N_k(s,a)\geq H} \frac{\sum_{h=1}^H \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{N_{k+1}(s,a)} + \sqrt{N_k(s,a)}} \\ &= 3 \sum_{k:N_k(s,a)\geq H} \left(\sqrt{N_{k+1}(s,a)} - \sqrt{N_k(s,a)}\right) \\ &\leq 3\sqrt{N_{K+1}(s,a)}\end{aligned} \tag{26}$$

where the inequality follows from (25), i.e., $4N_k(s,a) \geq 2N_k(s,a) \geq N_{k+1}(s,a)$, and the third equality follows from the fact that $\sum_{h=1}^H \mathbb{1}\{s_h^k = s, a_h^k = a\} = N_{k+1}(s,a) - N_k(s,a) = (\sqrt{N_{k+1}(s,a)} - \sqrt{N_k(s,a)})(\sqrt{N_{k+1}(s,a)} + \sqrt{N_k(s,a)})$. By applying (24) and (26) to (23), we have

$$\sum_{k=1}^K \frac{\sum_{h=1}^H \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{N_k(s,a) \vee 1}} \leq 2H + 3\sqrt{N_{K+1}(s,a)}.$$

By summing this over $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{k=1}^K \frac{\sum_{h=1}^H \mathbb{1}\{s_h^k = s, a_h^k = a\}}{\sqrt{N_k(s,a) \vee 1}} &\leq 2HSA + 3\sum_{s,a} \sqrt{N_{K+1}(s,a)} \\ &\leq 2HSA + 3\sqrt{SA}\sqrt{\sum_{s,a} N_{K+1}(s,a)} \\ &\leq 2HSA + 3\sqrt{SAT}.\end{aligned}$$

Finally, we have shown that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{2\beta}{\sqrt{N_k(s_h^k,a_h^k)\vee 1}} \le 2\beta\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\frac{\sum_{h=1}^{H}\mathbb{1}\{s_h^k=s,a_h^k=a\}}{\sqrt{N_k(s,a)\vee 1}} \le 4\beta HSA + 6\beta\sqrt{SAT}.$$

This concludes the proof. □

Lemma 16 (Formal statement of (6))**.** *Suppose that Lemmas 12 and 14 hold. Then we have*

$$\mathrm{Regret}_{aug}(T) \le 2C\sqrt{2T\log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT}.$$

Proof. For each $k\in[K]$, we have

$$V_{\tilde{r},1}^{\pi^*}(s_1^k,z_1^k) - \sum_{h=1}^{H}\tilde{r}(s_h^k,a_h^k,z_h^k) = \underbrace{V_{\tilde{r},1}^{\pi^*}(s_1^k,z_1^k) - V_{k,1}(s_1^k,z_1^k)}_{\text{(I)}} + \underbrace{V_{k,1}(s_1^k,z_1^k) - \sum_{h=1}^{H}\tilde{r}(s_h^k,a_h^k,z_h^k)}_{\text{(II)}}.$$

Note that (I) ≤ 0 by Lemma 13, and (II) can be rewritten as follows.

$$\begin{aligned}V_{k,1}(s_1^k,z_1^k) - \sum_{h=1}^{H}\tilde{r}(s_h^k,a_h^k,z_h^k) &= V_{k,1}(s_1^k,z_1^k) - \tilde{V}_{k,1}(s_1^k,z_1^k) + \tilde{V}_{k,1}(s_1^k,z_1^k) - Q_{k,1}(s_1^k,a_1^k,z_1^k)\\ &\quad + \left(\hat{P}_k(\cdot|s_1^k,a_1^k) - P(\cdot|s_1^k,a_1^k)\right)^\top V_{k,2}(\cdot,z_2^k)\\ &\quad + P(\cdot|s_1^k,a_1^k)^\top V_{k,2}(\cdot,z_2^k) - V_{k,2}(s_2^k,z_2^k)\\ &\quad + \beta/\sqrt{N_k(s_1^k,a_1^k)\vee 1}\\ &\quad + V_{k,2}(s_2^k,z_2^k) - \sum_{h=2}^{H}\tilde{r}(s_h^k,a_h^k,z_h^k).\end{aligned}$$

Each term can be bounded as follows. Due to the fact that $z_1^k\in\mathcal{Z}$ (Lemma 11) and the design of the clipping operation, we have $V_{k,1}(s_1^k,z_1^k) - \tilde{V}_{k,1}(s_1^k,z_1^k) \le 0$. Since $\tilde{V}_{k,1}(s_1^k,z_1^k) = \max_a Q_{k,1}(s_1^k,a,z_1^k) = Q_{k,1}(s_1^k,a_1^k,z_1^k)$, we have $\tilde{V}_{k,1}(s_1^k,z_1^k) - Q_{k,1}(s_1^k,a_1^k,z_1^k) = 0$. By Lemma 12, we have $(\hat{P}_k(\cdot|s_1^k,a_1^k) - P(\cdot|s_1^k,a_1^k))^\top V_{k,2}(\cdot,z_2^k) \le \beta/\sqrt{N_k(s_1^k,a_1^k)\vee 1}$. Applying these yields

$$\begin{aligned}V_{k,1}(s_1^k,z_1^k) - \sum_{h=1}^{H}\tilde{r}(s_h^k,a_h^k,z_h^k) &\le P(\cdot|s_1^k,a_1^k)^\top V_{k,2}(\cdot,z_2^k) - V_{k,2}(s_2^k,z_2^k)\\ &\quad + \frac{2\beta}{\sqrt{N_k(s_1^k,a_1^k)\vee 1}} + V_{k,2}(s_2^k,z_2^k) - \sum_{h=2}^{H}\tilde{r}(s_h^k,a_h^k,z_h^k).\end{aligned}$$

Due to the recursion, we have

$$\begin{aligned}\text{(II)} &= V_{k,1}(s_1^k,z_1^k) - \sum_{h=1}^{H}\tilde{r}(s_h^k,a_h^k,z_h^k)\\ &\le \sum_{h=1}^{H}\left(P(\cdot|s_h^k,a_h^k)^\top V_{k,h+1}(\cdot,z_{h+1}^k) - V_{k,h+1}(s_{h+1}^k,z_{h+1}^k)\right)\\ &\quad + \sum_{h=1}^{H}\frac{2\beta}{\sqrt{N_k(s_h^k,a_h^k)\vee 1}}.\end{aligned}$$

By applying bounds on (I) and (II), we have

$$V^{\pi^*}_{\tilde{r},1}(s^k_1, z^k_1) - \sum_{h=1}^{H} \tilde{r}(s^k_h, a^k_h, z^k_h) \le \sum_{h=1}^{H} \left(P(\cdot|s^k_h, a^k_h)^\top V_{k,h+1}(\cdot, z^k_{h+1}) - V_{k,h+1}(s^k_{h+1}, z^k_{h+1}) \right) + \sum_{h=1}^{H} \frac{2\beta}{\sqrt{N_k(s^k_h, a^k_h) \vee 1}}.$$

By summing this over $k = 1, \ldots, K$, we have

$$\begin{aligned}
&\sum_{k=1}^{K} \left(V^{\pi^*}_{\tilde{r},1}(s^k_1, z^k_1) - \sum_{h=1}^{H} \tilde{r}(s^k_h, a^k_h, z^k_h) \right) \\
&\le \sum_{k=1}^{K} \sum_{h=1}^{H} \left(P(\cdot|s^k_h, a^k_h)^\top V_{k,h+1}(\cdot, z^k_{h+1}) - V_{k,h+1}(s^k_{h+1}, z^k_{h+1}) \right) + \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{2\beta}{\sqrt{N_k(s^k_h, a^k_h) \vee 1}} \\
&\le 2C\sqrt{2T \log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT}.
\end{aligned}$$

□

C.4 Deferred proofs for Step 2 of Section 4

Lemma 17 (Formal statement of (7))**.** *Suppose that Lemmas 12 and 14 hold. Then we have*

$$\mathrm{Regret}(T) + \Phi_W(z_{T+1}) \le 2C\sqrt{2T \log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT} + KC.$$

Proof. Recall that $\tilde{r}(s, a, z) = r(s, a) + \Phi_W(z) - \Phi_W(\psi(s, a, z)$ and $\Phi_W(z^1_1) = 0$. Then we have

$$\begin{aligned}
-\sum_{t=1}^{T} \tilde{r}(s_t, a_t, z_t) &= -\sum_{t=1}^{T} \left(r(s_t, a_t) + \Phi_W(z_t) - \Phi_W(\psi(s_t, a_t, z_t)) \right) \\
&= -\sum_{t=1}^{T} \left(r(s_t, a_t) + \Phi_W(z_t) - \Phi_W(z_{t+1}) \right) \\
&= -\sum_{t=1}^{T} r(s_t, a_t) + \Phi_W(z_{T+1})
\end{aligned}$$

Then we have

$$\begin{aligned}
\mathrm{Regret}_{\text{aug}}(T) &= \sum_{k=1}^{K} \left(V^{\pi^*}_{\tilde{r},1}(s^k_1, z^k_1) - \sum_{h=1}^{H} \tilde{r}(s^k_h, a^k_h, z^k_h) \right) \\
&= TJ^*_r - \sum_{t=1}^{T} r(s_t, a_t) + \Phi_W(z_{T+1}) + \sum_{k=1}^{K} V^{\pi^*}_{\tilde{r},1}(s^k_1, a^k_1) - TJ^*_r \\
&= \mathrm{Regret}(T) + \Phi_W(z_{T+1}) + \sum_{k=1}^{K} V^{\pi^*}_{\tilde{r},1}(s^k_1, a^k_1) - TJ^*_r
\end{aligned}$$

Then it can be rewritten as

$$\mathrm{Regret}(T) + \Phi_W(z_{T+1}) = \mathrm{Regret}_{\text{aug}}(T) + \sum_{k=1}^{K} \left(HJ^*_r - V^{\pi^*}_{\tilde{r},1}(s^k_1, a^k_1) \right).$$

By Lemma 16, we have

$$\mathrm{Regret}_{\text{aug}}(T) \le 2C\sqrt{2T \log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT}.$$

By Lemma 10, we have

$$\sum_{k=1}^{K}\left(HJ_r^* - V_{\tilde{r},1}^{\pi^*}(s_1^k, a_1^k)\right) \le KC.$$

Finally, we have

$$\mathrm{Regret}(T) + \Phi_W(z_{T+1}) \le 2C\sqrt{2T\log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT} + KC.$$

□

C.5 Deferred proofs for Step 3 of Section 4

Lemma 18 (Formal statement of (8))**.** *Suppose that Lemmas 7, 12, and 14 hold. Then we have*

$$z_{T+1} \le \widetilde{\mathcal{O}}\left((\mathrm{sp}(v_r^*) + \mathrm{sp}(v_g^*) + \mathrm{sp}(v_g^*)^2)S^2A\sqrt{T}\right)$$

Proof. Recall that, by Lemmas 17 and 7, we have

$$\mathrm{Regret}(T) + \Phi_W(z_{T+1}) \le 2C\sqrt{2T\log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT} + KC,$$
$$\mathrm{Regret}(T) \ge -\lambda^* \,\mathrm{Violation}(T) - \mathrm{sp}(v_{\lambda^*}^*)\sqrt{2T\log(1/\delta)} - \mathrm{sp}(v_{\lambda^*}^*).$$

These lead to

$$\begin{aligned}\Phi_W(z_{T+1}) \le\ & 2C\sqrt{2T\log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT} + KC \\ & + \lambda^* \,\mathrm{Violation}(T) + \mathrm{sp}(v_{\lambda^*}^*)\sqrt{2T\log(1/\delta)} + \mathrm{sp}(v_{\lambda^*}^*)\end{aligned}$$

Moreover, by applying a similar argument in the proof of Lemma 11, we can show the following.

$$\begin{aligned}|\mathrm{Violation}(T) - z_{T+1}| &= \left|\sum_{t=1}^{T}(0 - g(s_t, a_t))) - z_{T+1}\right| \\ &= \left|-\sum_{t=1}^{T} g(s_t, a_t) - \sum_{t=1}^{T}(z_{t+1} - z_t)\right| \\ &= \left|\sum_{t=1}^{T}(z_t - g(s_t, a_t) - z_{t+1})\right| \\ &\le \sum_{t=1}^{T}|z_t - g(s_t, a_t) - z_{t+1}| \\ &\le \frac{\Delta T}{2}\end{aligned} \tag{27}$$

Thus, we have

$$\begin{aligned}\Phi_W(z_{T+1}) &\le 2C\sqrt{2T\log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT} + KC \\ &\quad + \lambda^* z_{T+1} + \lambda^* |\mathrm{Violation}(T) - z_{T+1}| \\ &\quad + \mathrm{sp}(v_{\lambda^*}^*)\sqrt{2T\log(1/\delta)} + \mathrm{sp}(v_{\lambda^*}^*) \\ &\le 2C\sqrt{2T\log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT} + KC \\ &\quad + \lambda^* z_{T+1} + \frac{\lambda^*\Delta T}{2} + \mathrm{sp}(v_{\lambda^*}^*)\sqrt{2T\log(1/\delta)} + \mathrm{sp}(v_{\lambda^*}^*).\end{aligned}$$

Then it can be rewritten as

$$\begin{aligned}\Phi_W(z_{T+1}) - \lambda^* z_{T+1} \le\ & 2C\sqrt{2T\log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT} + KC \\ & + \frac{\lambda^*\Delta T}{2} + \mathrm{sp}(v_{\lambda^*}^*)\sqrt{2T\log(1/\delta)} + \mathrm{sp}(v_{\lambda^*}^*).\end{aligned}$$

By the fourth statement of Lemma 8, we have $(\Lambda - \lambda^*)z_{T+1} - \Lambda W/2 \leq \Phi_W(z_{T+1}) - \lambda^* z_{T+1}$. Then it follows that

$$(\Lambda - \lambda^*)z_{T+1} - \Lambda W/2 \leq 2C\sqrt{2T\log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT} + KC + \frac{\lambda^*\Delta T}{2} + \mathrm{sp}(v^*_{\lambda^*})\sqrt{2T\log(1/\delta)} + \mathrm{sp}(v^*_{\lambda^*}).$$

Recall that $\Lambda = 2/\gamma$ and $\lambda^* \leq 1/\gamma$ by Lemma 6. Thus, we have $\Lambda - \lambda^* \geq 1/\gamma$. Then it follows that

$$z_{T+1} \leq 2\gamma C\sqrt{2T\log(2/\delta)} + 4\gamma\beta HSA + 6\gamma\beta\sqrt{SAT} + \gamma KC + \frac{\gamma\lambda^*\Delta T}{2} + \mathrm{sp}(v^*_{\lambda^*})\sqrt{2T\log(1/\delta)} + \gamma\,\mathrm{sp}(v^*_{\lambda^*}) + \frac{\gamma\Lambda W}{2}.$$

Again, recall that $\lambda^* \leq 1/\gamma$, $0 \leq \gamma \leq 1$, and the algorithmic parameter choice (5), i.e.,

$$\begin{aligned}
&\Delta = \frac{1}{\sqrt{T}},\ W = K = H = \sqrt{T},\\
&C = \mathrm{sp}(v^*_r) + \Lambda\,\mathrm{sp}(v^*_g) + \frac{\Lambda}{W}\left(\mathrm{sp}(v^*_g)^2 + 2H(\mathrm{sp}(v^*_g)+2)^2\right) + \frac{H\Lambda\Delta}{2},\\
&\quad = \mathcal{O}\left(\mathrm{sp}(v^*_r) + \frac{1}{\gamma}(\mathrm{sp}(v^*_g) + \mathrm{sp}(v^*_g)^2) + \frac{1}{\gamma}\right),\\
&\beta = 2CS\sqrt{\log(2TS^2|\mathcal{A}|/\delta)/2}\\
&\quad \leq \widetilde{\mathcal{O}}\left(S\,\mathrm{sp}(v^*_r) + \frac{S}{\gamma}(\mathrm{sp}(v^*_g) + \mathrm{sp}(v^*_g)^2) + \frac{S}{\gamma}\right).
\end{aligned}$$

Then it follows that

$$z_{T+1} \leq \widetilde{\mathcal{O}}\left((\mathrm{sp}(v^*_r) + \mathrm{sp}(v^*_g) + \mathrm{sp}(v^*_g)^2)S^2A\sqrt{T} + \mathrm{sp}(v^*_{\lambda^*})\sqrt{T}\right)$$

□

C.6 Proof of Theorem 2

By a union bound, Lemmas 7, 12, and 14 hold with probability at least $1 - 3\delta$. Throughout the proof, suppose that these lemmas hold. Then, by Lemma 17, we have

$$\mathrm{Regret}(T) + \Phi_W(z_{T+1}) \leq 2C\sqrt{2T\log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT} + KC. \tag{28}$$

For $\mathrm{Regret}(T)$, since $\Phi_W(z_{T+1}) \geq 0$, it directly follows that

$$\begin{aligned}
\mathrm{Regret}(T) &\leq 2C\sqrt{2T\log(2/\delta)} + 4\beta HSA + 6\beta\sqrt{SAT} + KC\\
&\leq \widetilde{\mathcal{O}}\left(\left(\mathrm{sp}(v^*_r) + \frac{1}{\gamma}(\mathrm{sp}(v^*_g) + \mathrm{sp}(v^*_g)^2 + 1)\right)S^2A\sqrt{T}\right)
\end{aligned}$$

Next, we bound $\mathrm{Violation}(T)$. By Lemma 18, we have

$$z_{T+1} \leq \widetilde{\mathcal{O}}\left((\mathrm{sp}(v^*_r) + \mathrm{sp}(v^*_g) + \mathrm{sp}(v^*_g)^2)S^2A\sqrt{T} + \mathrm{sp}(v^*_{\lambda^*})\sqrt{T}\right).$$

Moreover, by (27), we have $|\,\mathrm{Violation}(T) - z_{T+1}| \leq \Delta T/2 \leq \widetilde{\mathcal{O}}(\sqrt{T})$. Then it follows that

$$\begin{aligned}
\mathrm{Violation}(T) &\leq z_{T+1} + |\,\mathrm{Violation}(T) - z_{T+1}|\\
&\leq \widetilde{\mathcal{O}}\left((\mathrm{sp}(v^*_r) + \mathrm{sp}(v^*_g) + \mathrm{sp}(v^*_g)^2)S^2A\sqrt{T} + \mathrm{sp}(v^*_{\lambda^*})\sqrt{T}\right).
\end{aligned}$$

□

D Auxiliary Lemmas

Lemma 19. *Let $d \in \mathbb{Z}_+$, and let $p, q \in [0,1]^d$ such that $\sum_{i=1}^d p_i = \sum_{i=1}^d q_i = 1$. Let $v \in \mathbb{R}^d$. Then we have $|(p-q)^\top v| \leq \mathrm{sp}(v)||p-q||_1$.*

Proof. Let $v_{i_{\min}} = \arg\min_{j\in[d]} v_j$. Note that

$$\begin{aligned}
|(p-q)^\top v| &= \left|\sum_{i=1}^{d}(p_i-q_i)v_i\right| \\
&= \left|\sum_{i=1}^{d}(p_i-q_i)(v_i-v_{i_{\min}})\right| \\
&\le ||p-q||_1 \cdot \max_{j\in[d]}|v_j - v_{i_{\min}}| \\
&= ||p-q||_1 \cdot \mathrm{sp}(v)
\end{aligned}$$

where the second equality follows from the fact that $\sum_{i=1}^{d}(p_i-q_i)v_{i_{\min}} = 0$, the inequality follows from the Cauchy-Schwarz inequality, and the last equality follows from $\max_{j\in[d]}|v_j - v_{i_{\min}}| = \mathrm{sp}(v)$.

□

Lemma 20 (Lemma 4 in Chen et al. (2022))**.** *For any stationary policy π and reward function $r \in \mathbb{R}^{S\times|\mathcal{A}|}$ such that $J_r^\pi(s) = J_r^\pi$ for all $s \in \mathcal{S}$, we have $|V_{r,h}^\pi(s) - (H-h+1)J_r^\pi| \le \mathrm{sp}(v_r^\pi)$ for $(s,h) \in \mathcal{S}\times[H]$.*

E Experimental Details

Environment We use a tabular CMDP constructed by modifying the MDP of He et al. (2022) to incorporate a constraint and an infinite-horizon setting, following Yu et al. (2026). The state space is $\mathcal{S} = \{0,\dots,H_{\mathrm{env}}+1\}$, and the action space is $\mathcal{A} = \{-1,1\}^{m-1}$. For each state $s \in \{0,\dots,H_{\mathrm{env}}-1\}$ and action $a \in \mathcal{A}$, the constraint value is $g(s,a) = \mathrm{score}(a)$, where $\mathrm{score}(a)$ denotes the fraction of positive components in a. The process transitions to state $s+1$ with probability $0.95-0.01\cdot\mathbf{1}_{m-1}^\top a$ and to state $H_{\mathrm{env}}+1$ otherwise. State H_{env} yields reward 0 and state $H_{\mathrm{env}}+1$ yields reward 1. Both states incur zero constraint value and transition back to state 0, which makes the process infinite-horizon. The reward for $s \in \{0,\dots,H_{\mathrm{env}}-1\}$ is $r(s,a) = 0.4\cdot(1-\mathrm{score}(a))$ for $s < H_{\mathrm{env}}/2$ and $r(s,a) = 0.4\cdot\mathrm{score}(a)$ for $s \ge H_{\mathrm{env}}/2$, so in the early states the reward is given to the action that lowers the constraint value, whereas the CMDP of Yu et al. (2026) uses $r(s,a) = 0.4\cdot\mathrm{score}(a)$ in every state. We choose $H_{\mathrm{env}} = 6$, $m = 2$, $b = 0.7$, and $s_1 = 0$. We solve linear programs over occupancy measures to compute the optimal policy, J_r^*, and the Slater constant $\gamma = \max_\pi J_g^\pi - b$. We then compute the bias spans of the optimal policy.

Metrics and hyperparameters We report regret and constraint violation as

$$\mathrm{Regret}(t) = \sum_{\tau=1}^{t}(J_r^* - r(s_\tau,a_\tau)), \qquad \mathrm{Violation}(t) = \left[\sum_{\tau=1}^{t}(b - g(s_\tau,a_\tau))\right]_+.$$

Algorithm 1 is run with the shifted constraint function $g-b$. For Algorithm 1, we set $\Lambda = 2/\gamma$ and $H = W = \lfloor\sqrt{T}\rfloor$ as in Theorem 2, and $\Delta = 0.1/\lceil 0.1\sqrt{T}\rceil \le 1/\sqrt{T}$, for which the values of $g-b$ are integer multiples of Δ so that the projection Π_Δ has no rounding error. Following the adaptive clipping parameter $\zeta_k = \mathrm{sp}(v_r^*) + \lambda_k\,\mathrm{sp}(v_g^*)$ of Yu et al. (2026), we use the clipping parameter $C_k = \mathrm{sp}(v_r^*) + \Phi'_W(z_1^k)\,\mathrm{sp}(v_g^*)$, which replaces Λ in C by the slope of Φ_W at z_1^k, and set the bonus term to $\beta_k = C_k|\mathcal{S}|$, omitting the constant and the logarithmic factor. For the baselines, we use one-hot features with $d = |\mathcal{S}||\mathcal{A}| = 16$ and follow the parameter choices of Yu et al. (2026). That is, (i) for PD-LSCVI-UCB, we set $H = T^{1/3}$, the clipping parameter ζ_k, and $\beta_k = (\zeta_k+1)d$, and (ii) for Algorithm 2 of Ghosh et al. (2023), we follow its default parameters except that $H = \lceil T^{1/4}\rceil$ and $\beta = dH$. (iii) For Algorithm 3 of Chen et al. (2022), we set $H = \lceil(T/|\mathcal{S}|^2|\mathcal{A}|)^{1/3}\rceil = 7$ and solve its linear program over occupancy measures in every episode. Its constraint $\langle\nu,c\rangle \le H\tau + \mathrm{sp}(v_c^*)$ is written with the cost $c = 1-g$ and the threshold $\tau = 1-b$, which is equivalent to $\langle\nu,g\rangle \ge Hb - \mathrm{sp}(v_g^*)$. For its Bernstein-type confidence set, we use the radius $\sqrt{\bar{P}_k(s' \mid s,a)/N_k(s,a)} + 1/N_k(s,a)$ around the empirical transition $\bar{P}_k$, omitting the constants and the logarithmic factor, and the sensitivity experiment below multiplies this radius by c.

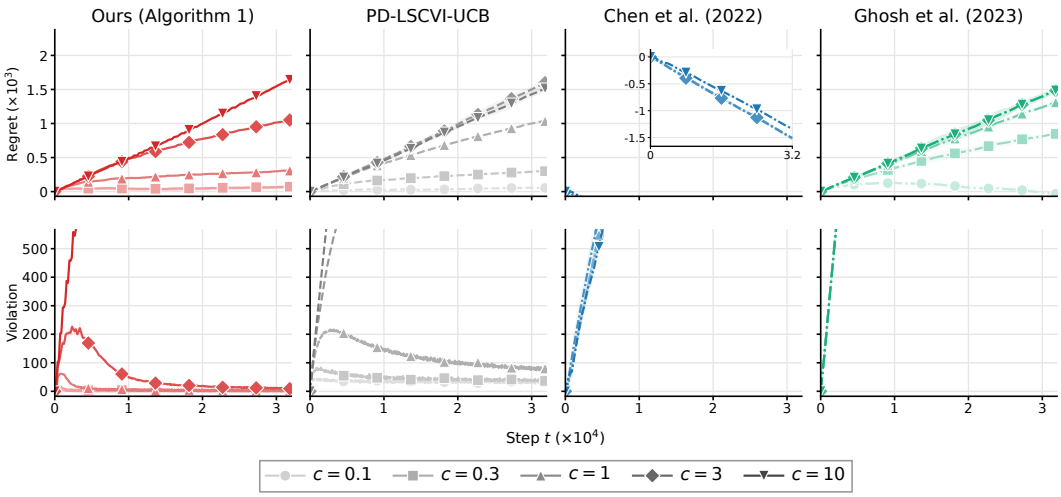


Figure 2: Plots of regret and constraint violation when the bonus term of each algorithm is multiplied by $c \in \{0.1, 0.3, 1, 3, 10\}$, where lighter curves correspond to smaller c. Each plot represents the average over 3 trials, and shaded regions indicate 95% confidence intervals.

Sensitivity to the bonus term The bonus terms above omit constants and logarithmic factors. To examine the sensitivity to their scale, we multiply the bonus term of each algorithm by $c \in \{0.1, 0.3, 1, 3, 10\}$, where $c = 1$ corresponds to Figure 1, and repeat 3 simulations. The results are summarized in Figure 2. Algorithm 1 achieves sublinear regret and constraint violation for every $c \leq 1$, its constraint violation is at most that of PD-LSCVI-UCB for every c, and its regret is smaller than or comparable to that of PD-LSCVI-UCB for every c. The regret of Algorithm 3 of Chen et al. (2022) is negative for every c, as shown in the inset, and its constraint violation is about $\mathrm{sp}(v_g^*)$ per episode for every c, since the slack in its linear program does not depend on c.

F Discussions and Future Work

Extension to linear CMDPs We discuss whether Algorithm 1 extends to average-reward linear CMDPs, where $P(s'|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}(s') \rangle$, $r(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_r \rangle$, and $g(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_g \rangle$ for a known feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ (Ghosh et al., 2023; Yu et al., 2026). Most of the analysis does not depend on the tabular structure. Lemmas 3 and 9 use only the Poisson equation of π^* and the properties of Φ_W. The violation analysis in the proof of Theorem 2 uses only strong duality. The empirical transition kernel can be replaced by least-squares value iteration for each debt value z. The sum of bonus terms in Lemma 15 can be bounded by the elliptical potential lemma (Abbasi-Yadkori et al., 2011). The value function clipping can be applied for each z as in Yu et al. (2026).

The difficulty lies in Lemma 12. In the tabular setting, the concentration of $\hat{P}_k$ holds for all bounded functions at once, and no covering argument is needed. In the linear setting, the regression error must be controlled uniformly over the class of value functions, and the bonus term scales with the square root of its log covering number (Jin et al., 2020). In Algorithm 1, the backup at (s,a,z) uses $V_{k,h+1}(\cdot, \Pi_\Delta(z - g(s,a)))$. The value function at a fixed z therefore depends on the regression weights of every debt value reachable from z in one step. When g is linear in the features, there are up to $2/\Delta$ such debt values, and the log covering number is $\mathcal{O}(d/\Delta + d^2)$ instead of $\mathcal{O}(d^2)$. Balancing the resulting bonus term $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T/\Delta})$ against the projection error $\mathcal{O}(\Delta T/\gamma)$ gives $\Delta \asymp T^{-1/3}$ and $\widetilde{\mathcal{O}}(T^{2/3})$ regret and constraint violation. This matches the bound of Yu et al. (2026) but not the $\widetilde{\mathcal{O}}(\sqrt{T})$ bound of Theorem 2. Designing an algorithm that resolves the growth of the covering number due to the reachable debt values is left for future work.

Limitations of the Primal-Dual Approach In Section 3, we discuss that extending the primal-dual approach to the average-reward setting is nontrivial. Here, we justify this by explaining the following two points: (i) the limitation of Yu et al. (2026)—a primal-dual algorithm for learn-

ing CMDPs—and (ii) the difficulty of extending Hong & Tewari (2025)—an algorithm for unconstrained setting—to the constrained setting.

For (i), Yu et al. (2026) proposes a primal-dual algorithm based on finite-horizon approximation that achieves only the suboptimal guarantee $\widetilde{\mathcal{O}}(T^{2/3})$. While their algorithm shares a similar structure with ours in that it also uses finite-horizon approximation, it incorporates the dual multiplier into the reward. In particular, the dual multiplier λ_{k+1} is updated in each episode k, which can be simplified as $\lambda_{k+1} = \lambda_k - \eta \sum_{h=1}^{H} g(s_h^k, a_h^k)$, where η is a step size. From the perspective of regret analysis, however, this update is problematic. Specifically, since $\sum_{h=1}^{H} g(s_h^k, a_h^k)$ can be as large as H, the standard dual-regret analysis incurs a term of the form $\lambda^2/(2\eta) + \eta K H^2/2$, which cannot be bounded by $\widetilde{\mathcal{O}}(\sqrt{T})$ under the relation $T = KH$. Consequently, their regret guarantee is limited to $\widetilde{\mathcal{O}}(T^{2/3})$.

For (ii), Hong & Tewari (2025) proposes a deviation-controlled value iteration algorithm for learning average-reward MDPs. A key step in their analysis is to control the deviation of value functions, whose analogue in our setting is $\|V_{k,h+1} - V_{k+1,h+1}\|_\infty$. However, when a dual multiplier λ_k is introduced, the algorithm operates with respect to the episode-dependent composite reward $r_k = r + \lambda_k g$. Consequently, even when the estimated transition kernel is fixed across consecutive episodes, the deviation satisfies a recursion of the form

$$\|V_{k,h+1} - V_{k+1,h+1}\|_\infty \approx \|r_{k+1} - r_k\|_\infty + \|\hat{P}_k(V_{k+1,h+2} - V_{k,h+2})\|_\infty.$$

Here, the additional variation term $\|r_{k+1} - r_k\|_\infty$, which is absent in the unconstrained setting, is induced by the primal-dual extension. Since $\lambda_{k+1} \neq \lambda_k$, this creates an additional obstacle to directly extending their deviation-control argument to the constrained setting.

Use of Generative-AI Tools

In this work, we used generative AI tools to obtain correct state augmentations via Huber potential, to find related work on state augmentation in reinforcement learning, and to assist with the implementation of numerical experiments. The potential-difference reward design, the drafting of the manuscript, and theoretical developments and proofs were carried out by the authors. We have carefully reviewed all AI-assisted work, including the experimental code. We take responsibility for the final content of this work, including text, claims, proofs, and artifacts produced with the aid of generative AI.