# Near-optimal Conservative Exploration in Reinforcement Learning under Episode-wise Constraints

**Donghao Li** [*1] **Ruiquan Huang** [*1] **Cong Shen** [2] **Jing Yang** [1]

## Abstract

This paper investigates conservative exploration in reinforcement learning where the performance of the learning agent is guaranteed to be above a certain threshold throughout the learning process. It focuses on the tabular episodic Markov Decision Process (MDP) setting that has finite states and actions. With the knowledge of an existing safe baseline policy, an algorithm termed as StepMix is proposed to balance the exploitation and exploration while ensuring that the conservative constraint is never violated in each episode with high probability. StepMix features a unique design of a mixture policy that adaptively and smoothly interpolates between the baseline policy and the optimistic policy. Theoretical analysis shows that StepMix achieves near-optimal regret order as in the constraint-free setting, indicating that obeying the stringent episode-wise conservative constraint does not compromise the learning performance. Besides, a randomization-based EpsMix algorithm is also proposed and shown to achieve the same performance as StepMix. The algorithm design and theoretical analysis are further extended to the setting where the baseline policy is not given a priori but must be learned from an offline dataset, and it is proved that similar conservative guarantee and regret can be achieved if the offline dataset is sufficiently large. Experiment results corroborate the theoretical analysis and demonstrate the effectiveness of the proposed conservative exploration strategies.

[*]Equal contribution [1]School of EECS, The Pennsylvania State University, University Park, PA, USA [2]ECE Department, University of Virginia, Charlottesville, VA, USA. Correspondence to: Jing Yang <yangjing@psu.edu>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

## 1. Introduction

One of the major obstacles that prevent state-of-the-art reinforcement learning (RL) algorithms from being deployed in real-world systems is the lack of performance guarantee throughout the learning process. In particular, for many practical systems, a reasonable albeit not necessarily optimal *baseline policy* is often in place, and RL is later brought in as a (supposedly) superior solution to replace the baseline. System designers want the potentially better RL policy, but are also wary of the possible performance degradation incurred by exploration during the learning process. This dilemma exists in many domains, including digital marketing, robotics, autonomous driving, healthcare, and networking; see Garcıa & Fernández (2015); Wu et al. (2016) for a detailed discussion of practical examples. It is desirable to have the RL algorithm perform nearly as well (or better) as the baseline policy *at all times*.

To address this challenge, *conservative exploration* has received increased interest in RL research over the past few years (Garcelon et al., 2020a; Yang et al., 2022; Efroni et al., 2020; Zheng & Ratliff, 2020; Xu et al., 2020; Liu et al., 2021). In the online learning setting, exploration of the unknown environment is necessary for RL to learn about the underlying Markov Decision Process (MDP). However, "free" exploration provides no guarantee on the RL performance, particularly in the early phases where the knowledge of the environment is minimal and the algorithm tends to explore almost randomly. To solve this problem, the vast majority of the conservative exploration literature relies on a key idea of invoking the baseline policy early on to build a conservative budget, which can be spent in later episodes to take explorative actions. This intuition, however, critically depends on the definition of the conservative constraint being the *cumulative* expected reward over a horizon falling below a certain threshold. If a more stringent constraint defined on a *per episode* basis is adopted, this idea becomes infeasible and it is unclear how conservative exploration can be achieved.

In this paper, we focus on conservative exploration in an episodic MDP with finite states and actions. Unlike most of the prior works, we enforce a more strict conservative constraint that the expected reward of the RL policy cannot

*Table 1.* Comparison of Related Algorithms

| Algorithm (Reference) | Regret | Violation | Constraint-type | Baseline Assumption |
|---|---|---|---|---|
| BPI-UCBVI (Ménard et al., 2021) | $\tilde{O}\left(\sqrt{H^3SAK}\right)$ | N/A | N/A | N/A |
| OptPess-LP (Liu et al., 2021) | $\tilde{O}(\frac{1}{\kappa}\sqrt{H^6S^3AK})$ | 0 | Episodic, general constraint | Type I |
| DOPE (Bura et al., 2022) | $\tilde{O}(\frac{1}{\kappa}\sqrt{H^6S^2AK})$ | 0 | Episodic, general constraint | Type I |
| Budget-Exporation (Yang et al., 2022) | $\tilde{O}\left(\sqrt{H^3SAK}+\frac{H^3SA\Delta_0}{\kappa(\kappa+\Delta_0)}\right)$ | 0 | Cumulative, conservative constraint | Type I |
| StepMix / EpsMix (this work) | $\tilde{O}\left(\sqrt{H^3SAK}+\frac{H^3SA\Delta_0}{\kappa^2}\right)$ | 0 | Episodic, conservative constraint | Type I and II |
| Lower bound (Yang et al., 2022) | $\Omega\left(\sqrt{H^3SAK}+\frac{H^3SA\Delta_0}{\kappa(\kappa+\Delta_0)}\right)$ | 0 | Cumulative, conservative constraint | Type I |

$\Delta_0$: suboptimality gap for the baseline policy; $\kappa$: tolerable reward loss from the baseline policy or the Slater parameter. Type I assumes a known safe baseline policy. Type II assumes availability of an offline dataset generated by an unknown safe behavior policy. The lower bound automatically applies to our problem, due to its weaker constraint.

be much worse than that of a baseline policy *for every episode*. One fundamental question we aim to answer is:

*Is it possible to design a conservative exploration algorithm to achieve the optimal learning regret while satisfying the episode-wise conservative constraint throughout the learning process?*

In this work, we provide an affirmative answer to this question. Our main contributions are summarized as follows.

- First, we investigate the scenario where a safe baseline policy is explicitly given upfront, and propose a model-based learning algorithm coined StepMix. In contrast to conventional linear programming or primal-dual based approaches in constrained MDPs (Liu et al., 2021; Bura et al., 2022; Wei et al., 2022; Efroni et al., 2020), StepMix features several unique design components. First, in order to achieve the optimal learning regret, StepMix relies on a Bernstein inequality-based design to closely track the estimation uncertainty in learning and construct an efficient optimistic policy correspondingly. Then, a set of candidate policies are explicitly constructed by smoothly interpolating between the safe baseline policy and the optimistic policy. Finally, a mixture of two candidate policies is obtained when necessary to achieve the near-optimal tradeoff between safe exploration and efficient exploitation.
- Second, we theoretically analyze the performance of StepMix, and rigorously show that it achieves $\tilde{O}(\sqrt{H^3SAK})$ regret, which is the order-optimal learning regret in the unconstrained setting, while never violating the conservative constraint during the learning process with high probability. The conservative constraint turns out to only incur an *additive* regret term, as opposed to a multiplicative coefficient in Bura et al. (2022); Liu et al. (2021). Furthermore, the additive term differs from that in the lower bound in Yang et al. (2022) by a small constant factor, while our constraint is more stringent. Besides, we extend the analysis to a randomization mechanism-based EpsMix algorithm and show that it achieves the same learning regret as StepMix and satisfies the conservative constraint as well. A comparison of our work and these relevant papers is presented in Table 1.
- Next, instead of assuming a safe baseline policy is explicitly provided, we investigate the scenario where the agent only has access to an offline dataset collected under an unknown safe behavior policy. The agent thus needs to first extract an approximately safe baseline policy from the dataset and then to use it as an input to the StepMix or EpsMix algorithm. We explicitly characterize the impact of the dataset size and the quality of the behavior policy on the safety and regret of StepMix/EpsMix. Our results indicate that similar regret and safety guarantees can be achieved, as long as the dataset is sufficiently large.
- Finally, due to the explicit algorithmic design of the optimistic policy, the candidate policies and the mixture policies, we are able to implement StepMix and EpsMix efficiently and validate their performances through synthetic experiments. The experimental results corroborate our theoretical findings, and showcase the superior performances of StepMix/EpsMix compared with other baseline algorithms.

## 2. Related Works

In this section, we briefly discuss existing works that are most relevant to our work. A detailed literature review is deferred to Appendix A.

**Unconstrained Episodic Tabular MDPs.** Unconstrained tabular MDPs have been well studied in the literature. For an episodic MDP with $S$ states, $A$ actions and horizon $H$, the minimax regret lower bound scales in $\Omega(\sqrt{H^3SAK})$ (Domingues et al., 2021a), where $K$ denotes the number of episodes. Several algorithms have been proposed and shown to achieve the minimax lower bound (and thus order-optimal), including Azar et al. (2017); Zanette & Brunskill (2019); Ménard et al. (2021).

**Conservative Exploration.** Conservative exploration corresponds to the setting where a good baseline policy that may not be optimal is available, and the agent is required to perform not much worse than the baseline policy during the learning process. Such conservative scenario has been studied in bandits (Wu et al., 2016; Kazerouni et al., 2017; Garcelon et al., 2020b) and tabular MDPs (Garcelon et al., 2020a). Garcelon et al. (2020a) investigate both the average reward setting and the finite horizon setting. Yang et al. (2022) propose a reduction-based framework for conservative bandits and RL, which translates a minimax lower bound of the non-conservative setting to a valid lower bound for the conservative case. It also proposes a Budget-Exporation algorithm and shows that its regret scales in $\tilde{O}\left(\sqrt{H^3SAK} + \frac{H^3SA\Delta_0}{\kappa(\kappa+\Delta_0)}\right)$ for tabular MDPs, where $\Delta_0$ is the suboptimality gap of the baseline policy, and $\kappa$ is the tolerable performance loss from the baseline. However, all these works assume *cumulative* conservative constraint. As discussed in Section 1, our episodic-wise constraint is more stringent, and correspondingly the algorithms and the regret analysis are also different from the prior works.

**Constrained MDP with Baseline Policies.** Conservative exploration studied in this paper can be viewed as a specific case of the Constrained Markov Decision Process (CMDP) (Altman, 1999; Liu et al., 2021; Efroni et al., 2020; Wei et al., 2022), where the goal is to maximize the expected total reward subject to constraints on the expected total costs in each episode. Assuming a known safe baseline policy that satisfies the corresponding constraints, OptPess-LP (Liu et al., 2021) is shown to achieve an regret of $\tilde{O}(\frac{1}{\kappa}\sqrt{H^6S^3AK})$ without any constraint violation with high probability, while DOPE (Bura et al., 2022) improves the regret to $\tilde{O}(\frac{1}{\kappa}\sqrt{H^6S^2AK})$, where $\kappa$ denotes the Slater parameter. We note that both algorithms do not achieve the optimal regret in the unconstrained counterpart.

## 3. Problem Formulation

We consider an episodic MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r, s_1)$, where $\mathcal{S}$ and $\mathcal{A}$ are the sets of states and actions, respectively, $H \in \mathbb{Z}_+$ is the length of each episode, $P = \{P_h\}_{h=1}^H$ and $r = \{r_h\}_{h=1}^H$ are respectively the state transition probability measures and the reward functions, and $s_1$ is a given initial state. We assume that $\mathcal{S}$ and $\mathcal{A}$ are finite sets with cardinality $S$ and $A$ respectively. Moreover, for each $h \in [H]$, $P_h(\cdot|s,a)$ denotes the transition kernel over the next states if action $a$ is taken for state $s$ at step $h \in [H]$, and $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the deterministic reward function at step $h$ which is assumed be known for simplicity. Our result can be easily generalized to random and unknown reward functions. We consider the learning problem where $\mathcal{S}$ and $\mathcal{A}$ are known while $P$ are unknown a priori.

A policy $\pi$ is a set of mappings $\{\pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}_{h\in[H]}$, where $\Delta(\mathcal{A})$ is the set of all probability distributions over the action space $\mathcal{A}$. In particular, $\pi_h(a|s)$ denotes the probability of selecting action $a$ in state $s$ at time step $h$.

An agent interacts with this episodic MDP as follows. In each episode, the environment begins with a fixed initial state $s_1$. Then, at each step $h \in [H]$, the agent observes the state $s_h \in \mathcal{S}$, picks an action $a_h \in \mathcal{A}$, and receives a reward $r_h(s_h, a_h) \in [0,1]$. The MDP then evolves to a new state $s_{h+1}$ that is drawn from the probability measure $P_h(\cdot|s_h, a_h)$. The episode terminates after $H$ steps.

For each $h \in [H]$, we define the state-value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ as the expected total reward received under policy $\pi$ when starting from an arbitrary state at the $h$-th step until the end of the episode. Specifically, $\forall s \in \mathcal{S}, h \in [H]$,

$$V_h^\pi(s) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'})\middle| s_h = s\right], \tag{1}$$

where we use $\mathbb{E}_\pi[\cdot]$ to denote the expectation over states and actions that are governed by $\pi$ and $P$. Since the MDP begins with the same initial state $s_1$, to simplify the notation, we use $V^\pi$ to denote $V_1^\pi(s_1)$ without causing ambiguity. Correspondingly, we define the action-value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ at step $h$ as the expected total reward under policy $\pi$ after taking action $a$ at state $s$ in step $h$, that is:

$$\begin{aligned} Q_h^\pi(s,a) :=& \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'})\middle| s_h = s, a_h = a\right] \\ =& r_h(s,a) + [P_h V_{h+1}^\pi](s,a), \end{aligned}$$

where $[P_h V_{h+1}^\pi](s,a) := \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}^\pi(s')]$. Since the action space and the episode length are both finite, there always exists an optimal policy $\pi^\star$ that gives the optimal value $V_h^\star(s) = \sup_\pi V_h^\pi(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$.

**Conservative Constraint.** While there could be various forms of constraints imposed on the RL algorithms, in this work, we focus on a baseline policy-based constraint (Garcelon et al., 2020c; Yang et al., 2022). In many applications, it is common to have a known and reliable baseline policy that is potentially suboptimal but satisfactory to some degree. Therefore, for applications of RL algorithms, it is important that they are guaranteed to perform not much worse than the existing baseline throughout the learning process. Denote the baseline policy as $\pi^b$ and the corresponding expected total reward obtained under $\pi^b$ in an episode as $V_1^{\pi^b}$. Then, throughout the entire learning process, we require that the expected total reward for each episode $k$ is at least $\gamma$ with high probability, where $\kappa := V_1^{\pi^b} - \gamma > 0$ characterizes how much risk the algorithm can take during the learning process. A policy $\pi$ that achieves expected total reward at least $\gamma$ is considered to

be "safe", and we emphasize that our proposed algorithms do not require the knowledge of $V_1^{\pi^b}$. Let $\pi^k$ be the policy adopted by the agent during episode $k \in [K]$. Mathematically, we formulate the conservative constraint as

$$\mathbb{P}\left[V_1^{\pi^k} \geq \gamma, \forall k \in [K]\right] \geq 1-\delta, \text{ where } \delta \in (0,1). \quad (2)$$

**Comparison with Previous Conservative Constraints.** The conservative constraint in Equation (2) is more restrictive compared with Garcelon et al. (2020c); Yang et al. (2022), where the constraint is imposed on the cumulative expected reward over all experienced episodes instead of on each episode. We note that this stringent constraint has a profound impact on the algorithm design. While the previous cumulative conservative constraint enables the idea of saving the conservative budget early on and spending it later to play explorative actions, it cannot guarantee that in each episode, the expected total reward is above a certain threshold. Our constraint in Equation (2), in contrast, requires the expected total reward to be above a threshold in each episode. Hence, the idea of saving budget from early episodes for exploration in future episodes cannot be adopted, and it requires a more sophisticated algorithm design to control the budget spending *within* each episode and ensure the safety of all executed policies.

In addition, the per-episode conservative constraint in our work is more practical than the cumulative reward-based constraints. This is because each episode in the episodic MDP setting corresponds to the learning agent interacting with the environment from the beginning to the end, e.g., a robot walks from a starting point to the end point. Guaranteeing the performance in every episode has physical meanings, e.g., making sure that the robot does not suffer any damage while learning how to walk. This cannot be captured by the long-term constraint that spans many episodes.

**Learning Objective.** Under the given episodic MDP setting, the agent aims to learn the optimal policy by interacting with the environment during a set of episodes, subject to the conservative constraint. The difference between $V_1^{\pi^k}$ and $V_1^\star$ serves as the expected regret or the suboptimality of the agent in the $k$-th episode. Thus, after playing for $K$ episodes, the total expected regret is

$$\text{Reg}(K) := KV_1^\star - \sum_{k=1}^{K} V_1^{\pi^k}. \quad (3)$$

Our objective is to minimize $\text{Reg}(K)$ while satisfying Equation (2) for any given $\delta \in (0,1)$.

## 4. The StepMix Algorithm

In this section, we aim to design a novel safe exploration algorithm to satisfy the episodic conservative constraint and achieve the optimal learning regret.

### 4.1. Challenges

For unconstrained episodic MDPs with finite states and actions, in order to achieve the minimax regret lower bound $\Omega(\sqrt{H^3SAK})$, the core design principle (Azar et al., 2017; Zanette & Brunskill, 2019; Ménard et al., 2021) is to construct a Bernstein inequality-based Upper Confidence Bound (UCB) for the action-state value function under the optimal policy (i.e., $Q_h^\star$), and then to execute an optimistic policy that maximizes the UCB in each step. Such a UCB takes the variance of the corresponding estimated value function into consideration, leading to a more efficient exploration policy.

Intuitively, in order to achieve the same learning regret, the safe exploration policy should follow a similar Bernstein-inequality based design principle. However, this may lead to several technical challenges, as elaborated below.

First, we note that in conventional CMDP problems under *episodic* cost constraints (Bura et al., 2022; Liu et al., 2021), the exploration policy in each episode $k$ is usually obtained by solving a constrained optimization problem in the form of $\pi^k = \arg\max_{\pi \in \Pi_k} V^\pi(P^k)$, where $P^k$ is the estimated model and $\Pi_k$ is the set of estimated safe policies. For given $P^k$, both the objective function and the constraint set can be expressed as a linear function of $\pi$ or of occupancy measures, and thus can be solved efficiently. However, if Bernstein inequality is adopted to construct a tighter confidence set of the value functions (and hence $\Pi_k$), it can no longer be formulated as a linear programming problem, resulting in unfavorable computational complexity in each iteration.

Second, in order to keep track of the estimation error under the adopted exploration policy $\pi^k$, it is necessary to bound $(\hat{P}_h^k - P_h)V_{h+1}^{\pi^k}$ for each $h \in [H]$. In order to achieve the optimal dependence on $S$, a common technique is to decompose it into two terms, $(\hat{P}_h^k - P_h)V_{h+1}^{\pi^\circ}$ and $(\hat{P}_h^k - P_h)(V_{h+1}^{\pi^k} - V_{h+1}^{\pi^\circ})$, where $\pi^\circ$ is a fixed policy that is independent with the historical data, and then bound them separately. Intuitively, $\pi^\circ$ should be a policy "close" to $\pi^k$, so that as $\hat{P}_h^k$ converges to $P_h$, both terms converge to zero and the overall learning regret can thus be bounded. In the unconstrained case, $\pi^\circ$ is naturally set to be the optimal policy $\pi^\star$. However, under the episodic conservative constraint in our setting, the selection of $\pi^\circ$ is more delicate. This is because $\pi^k$ may be very different from $\pi^\star$, especially at the beginning of the learning stage when little information of the underlying MDP is known. Therefore, how to construct a good "anchor" policy $\pi^\circ$ that stays close to $\pi^k$ *throughout the learning process* becomes challenging.

Finally, in order to ensure the safety of the exploration pol-

icy $\pi_k$ in each episode, it is necessary to obtain a *pessimistic* estimation of the corresponding value function and to make sure it is above the threshold $\gamma$. While the Lower Confidence Bound (LCB) under the optimal policy $\pi^\star$ can be constructed in a symmetric manner as UCB, it is not immediately clear how to construct a Bernstein-type LCB for $V^{\pi_k}$, as $\pi_k$ is not fixed but dependent on history, and it may deviate from the optimistic policy significantly due to the episodic conservative constraint.

### 4.2. Algorithm Design

In this subsection, we explicitly address the aforementioned challenges and present a novel algorithm termed as StepMix. Before we proceed to elaborate the design of StepMix, we first introduce the definition of step mixture policies.

**Definition 4.1** (Step Mixture Policies)**.** The step mixture policy of two Markov policies $\pi^1$ and $\pi^2$ with parameter $\rho$, denoted by $\rho\pi^1+(1-\rho)\pi^2$, is a Markov policy such that the probability of choosing an action $a_h$ given a state $s_h$ under the step mixture policy is $\rho\pi_h^1(a_h|s_h)+(1-\rho)\pi_h^2(a_h|s_h)$.

StepMix is a model-based algorithm that features a unique design of the candidate policies and safe exploration policies. In the following, we elaborate its major components.

**Model Estimation.** At each episode $k$, the agent uses the available dataset to obtain an estimate of the transition kernel. Specifically, let $n_h^k(s,a)=\sum_{\tau=1}^{k-1}\mathbb{1}\{s_h^\tau=s,\ a_h^\tau=a\}$ and $n_h^k(s,a,s')=\sum_{\tau=1}^{k-1}\mathbb{1}\{s_h^\tau=s,\ a_h^\tau=a,\ s_{h+1}^\tau=s'\}$ be the visitation counters. The agent estimates $\hat{P}_h^k(s'|s,a)$ as

$$\hat{P}_h^k(s'|s,a)=\begin{cases}\frac{n_h^k(s,a,s')}{n_h^k(s,a)}, & \text{if } n_h^k(s,a)>1,\\ \frac{1}{S}, & \text{otherwise.}\end{cases} \tag{4}$$

**Bernstein-type Optimistic Policy Identification.** With the updated model estimates $\hat{P}^k$, the agent then tries to construct an optimistic policy. We note that this optimistic policy may not be identical to the exploration policy selected afterwards. However, it provides important information regarding the model estimate accuracy and will be leveraged to construct an efficient yet safe exploration policy. Specifically, we first denote

$$\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^k)(s,a)=\mathbb{E}_{s'\sim\hat{P}_h^k(\cdot|s,a)}[(\tilde{V}_{h+1}^k(s')-\mathbb{E}_{s'\sim\hat{P}_h^k(\cdot|s,a)}[\tilde{V}_{h+1}^k(s')])^2],$$

which captures the variance of $\tilde{V}_{h+1}^k$ under transition kernel $\hat{P}_h^k$ given $(s_h^k,a_h^k)=(s,a)$.

Then, with $\tilde{V}_{H+1}^k(s)=\underset{\sim}{V}_{H+1}^k(s)=0,\ \forall s\in\mathcal{S}$, for each $h\in[H]$, $(s,a)\in\mathcal{S}\times\mathcal{A}$, we recursively define

$$\tilde{Q}_h^k(s,a)\triangleq\min\Big(H,r_h(s,a)+3\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^k)(s,a)\frac{\beta^\star}{n_h^k(s,a)}}$$
$$+14H^2\frac{\beta}{n_h^k(s,a)}+\frac{1}{H}\hat{P}_h^k(\tilde{V}_{h+1}^k-\underset{\sim}{V}_{h+1}^k)(s,a)+\hat{P}_h^k\tilde{V}_{h+1}^k(s,a)\Big)$$

**Algorithm 1** The StepMix Algorithm

**Input:** $\pi^b$, $\gamma$, $\beta$, $\beta^\star$, $\mathcal{D}_0=\emptyset$.
**for** $k=1$ to $K$ **do**
  Update model estimate $\hat{P}$ according to Equation (4).
  *# Optimistic policy identification*
  $\tilde{V}_{H+1}^k(s)=\underset{\sim}{V}_{H+1}^k(s)=0,\forall s\in\mathcal{S}$.
  **for** $h=H$ to $1$ **do**
    Update $\tilde{Q}_h^k(s,a),\underset{\sim}{Q}_h^k(s,a),\forall(s,a)\in\mathcal{S}\times\mathcal{A}$ according to Equation (5).
    $\bar{\pi}_h^k(s)\leftarrow\arg\max_a\tilde{Q}_h^k(s,a)$, $\tilde{V}_h^k(s)\leftarrow\tilde{Q}_h^k(s,\bar{\pi}_h^k(s))$, $\underset{\sim}{V}_h^k(s)\leftarrow\underset{\sim}{Q}_h^k(s,\bar{\pi}_h^k(s)),\forall s\in\mathcal{S}$.
  **end for**
  *# Candidate policy construction and evaluation*
  **for** $h_0=0$ to $H$ **do**
    $\pi^{k,h_0}=\{\pi_1^b,\pi_2^b,\cdots,\pi_{h_0}^b,\bar{\pi}_{h_0+1}^k,\cdots,\bar{\pi}_{H-1}^k,\bar{\pi}_H^k\}$.
    $\underset{\sim}{V}^{k,h_0}=\mathrm{PolicyEva}(\hat{P}^k,\pi^{k,h_0})$.
  **end for**
  *# Safe exploration policy selection*
  **if** $\{h\,|\,\underset{\sim}{V}_1^{k,h}\geq\gamma,h=0,1,\ldots,H\}=\emptyset$ **then**
    $\pi^k=\pi^b$.
  **else**
    $h^k=\min\{h\,|\,\underset{\sim}{V}_1^{k,h}\geq\gamma,h=0,1,\ldots,H\}$.
    **if** $h^k=0$ **then**
      $\pi^k=\bar{\pi}^k$.
    **else**
      Set $\pi^k$ according to Equation (7).
    **end if**
  **end if**
  Execute $\pi^k$ and collect $\{(s_h^k,a_h^k,s_{h+1}^k)\}_{h=1}^H$.
  $\mathcal{D}_n\leftarrow\mathcal{D}_{n-1}\cup\{(s_h^k,a_h^k,s_{h+1}^k)\}_{h=1}^H$.
**end for**

$$\underset{\sim}{Q}_h^k(s,a)\triangleq\max\Big(0,r_h(s,a)-3\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^k)(s,a)\frac{\beta^\star}{n_h^k(s,a)}}$$
$$-22H^2\frac{\beta}{n_h^k(s,a)}-\frac{2}{H}\hat{P}_h^k(\tilde{V}_{h+1}^k-\underset{\sim}{V}_{h+1}^k)(s,a)+\hat{P}_h^k\underset{\sim}{V}_{h+1}^k(s,a)\Big), \tag{5}$$

and obtain an optimistic policy $\bar{\pi}^k$ by setting $\bar{\pi}_h^k(s)=\arg\max_{a\in\mathcal{A}}\tilde{Q}_h^k(s,a)$. After that, we set $\tilde{V}_h^k(s)=\tilde{Q}_h^k(s,\bar{\pi}_h^k(s))$, and $\underset{\sim}{V}_h^k(s)=\underset{\sim}{Q}_h^k(s,\bar{\pi}_h^k(s))$.

Intuitively speaking, $\tilde{Q}_h^k(s,a)$ serves as a Bernstein-type UCB for the true value function under the optimal policy, i.e., $Q^\star(s,a)$, while $\underset{\sim}{Q}_h^k(s,a)$ serves as the corresponding LCB. We note that the designs of $\tilde{Q}_h^k(s,a)$ and $\underset{\sim}{Q}_h^k(s,a)$ are not symmetric, i.e., the coefficients associated with the Bernstein-type bonus terms are not exactly opposite. Actually, this unique selection of the bonus terms is critical for us to obtain a valid LCB not just for the optimal value function, but also for those value functions under the exploration policy $\pi^k$, as elaborated below.

**Candidate Policy Construction and Evaluation.** Once the agent obtains the optimistic policy $\bar{\pi}^k$, it will proceed to

construct a set of candidate policies denoted as $\{\pi^{k,h_0}\}_{h_0=0}^{H}$, where $\pi^{k,h_0} = \{\pi_1^b, \cdots \pi_{h_0}^b, \bar{\pi}_{h_0+1}^k, \cdots, \bar{\pi}^k\}$. We note that $\pi^{k,h_0}$ follows the safe baseline $\pi^b$ for the first $h_0$ steps, after which it switches to the optimistic policy $\bar{\pi}^k$. Besides, $\pi^{k,h_0}$ and $\pi^{k,h_0+1}$ only differ at step $h_0+1$. As $h_0$ sweeps from $H$ to 0, $\pi^{k,h_0}$ essentially forms a smooth interpolation between the safe baseline $\pi^b$ and the optimistic policy $\bar{\pi}^k$.

For each candidate policy $\pi^{k,h_0}$, we obtain UCB and LCB on the two corresponding true value functions denoted as $Q^{k,h_0}$ and $V^{k,h_0}$ respectively, by invoking the PolicyEva subroutine (See Algorithm 3 in Appendix C). Specifically, PolicyEva recursively updates $\tilde{Q}_h^{k,h_0}(s,a)$ and $\underset{\sim}{Q}_h^{k,h_0}(s,a)$ in the same form as in Equation (5), while $\tilde{V}_h^{k,h_0}(s) \triangleq \langle \pi_h^{k,h_0}(\cdot|s), \tilde{Q}_h^{k,h_0}(s,\cdot)\rangle$, $\underset{\sim}{V}_h^{k,h_0}(s) \triangleq \langle \pi_h^{k,h_0}(\cdot|s), \underset{\sim}{Q}_h^{k,h_0}(s,\cdot)\rangle$.

We design the set of candidate policies in order to explicitly address the second challenge in Section 4.1, i.e., it is desirable to obtain a fixed "anchor" policy that stays close to $\pi^k$ throughout the learning process. Intuitively, in order to satisfy the conservative constraint in each episode, $\pi^k$ would stay at $\pi^b$ when it has not collected enough information of the environment; As $k$ proceeds, it is desirable to have $\pi^k$ evolve to the optimal policy $\pi^\star$, in order to achieve the optimal learning regret. Thus, it may not be reasonable to expect that a single fixed anchor policy would stay close to $\pi^k$ in every episode. Instead, we construct *a set* of anchor policies denoted as $\{\pi^{\star,h_0}\}_{h_0=0}^{H}$, where $\pi^{\star,h_0} = \{\pi_1^b, \cdots \pi_{h_0}^b, \pi_{h_0+1}^\star, \cdots, \pi_H^\star\}$. Essentially, $\pi^{k,h_0}$ is the optimistic version of $\pi^{\star,h_0}$. Thus, the estimation error in $(\hat{P}_h^k - P_h)V_{h+1}^{\pi^{k,h_0}}$ can be decomposed with respect to $V_{h+1}^{\pi^{\star,h_0}}$ and then be bounded separately. As $\pi^k$ dynamically evolves in between $\pi^b$ and $\bar{\pi}$, we expect that it stays close to $\pi^{k,h_0}$ for certain $h_0$, and thus the estimation error in $(\hat{P}_h^k - P_h)V_{h+1}^{\pi^k}$ can be effectively bounded as well.

**Safe Exploration Policy Selection.** After constructing and evaluating the set of candidate policies, we then design a safe exploration policy by mixing two neighboring candidate policies.

Specifically, the learner will compare $\underset{\sim}{V}_1^{k,h_0}$ with the threshold $\gamma$ for $h_0 = 0, 1, \ldots, H$. If it is above the threshold, it indicates that with high probability the candidate policy $\pi^{k,h_0}$ will satisfy the conservative constraint. Let $h^k$ be the smallest $h_0$ such that $\underset{\sim}{V}_1^{k,h_0} \geq \gamma$. Then, we have the following cases:

- If $h^k = 0$, it indicates that the LCB of the optimistic policy $\bar{\pi}$ is above the threshold. Thus the learner executes $\bar{\pi}^k$.
- If $h^k \in [1:H]$, it indicates that $\pi^{k,h^k}$ is safe but $\pi^{k,h^k-1}$ may be not. More importantly, they only differ in a single step $h^k$. Then, the learner would construct a mixture of $\pi^{k,h^k}$ and $\pi^{k,h^k-1}$ as follows:

$$\rho = \frac{\underset{\sim}{V}_1^{k,h^k}(s_1) - \gamma}{\underset{\sim}{V}_1^{k,h^k}(s_1) - \underset{\sim}{V}_1^{k,h^k-1}(s_1)}, \tag{6}$$

$$\pi^k = (1-\rho)\pi^{k,h^k} + \rho\pi^{k,h^k-1}. \tag{7}$$

- If none of $\underset{\sim}{V}^{k,h_0}$ is above the threshold, it indicates that the LCB of $V^{\pi^b}$ is below the threshold, which occurs when the estimation has high uncertainty. The learner will then resort to $\pi^b$ for conservative exploration.

Once policy $\pi^k$ is executed and a trajectory is collected, the learner moves on to the next episode.

### 4.3. Theoretical Analysis

The performance of StepMix is stated in the following theorem.

**Theorem 4.2** (Informal). *With probability at least $1-\delta$, StepMix (Algorithm 1) simultaneously (i) satisfies the conservative constraint in Equation (2), and (ii) achieves a total regret that is at most*

$$\tilde{O}\big(\sqrt{H^3SAK} + H^3S^2A + H^3SA\Delta_0\big(\tfrac{1}{\kappa^2} + \tfrac{S}{\kappa}\big)\big),$$

*where $\Delta_0 := V_1^\star - V_1^{\pi^b}$ is the suboptimality gap of the baseline policy and $\kappa := V_1^{\pi^b} - \gamma$ is the tolerable value loss from the baseline policy.*

*Remark* 4.3. Theorem 4.2 indicates that StepMix achieves a near-optimal regret in the order of $\tilde{O}(\sqrt{H^3SAK})$, while ensuring zero constraint violation with high probability. Compared with BPI-UCBVI (Ménard et al., 2021), the conservative exploration only leads to an additive constant term $\tilde{O}(H^3SA\Delta_0(\frac{1}{\kappa^2} + \frac{S}{\kappa}))$ in the learning regret bound. The additive term matches with that in the lower bound under the weaker cumulative conservative constraint in Yang et al. (2022) up to a constant, indicating our result is near-optimal. For the special case when $\gamma = 0$, the LCBs estimated in StepMix will always be greater than $\gamma$; thus the optimistic policy is always safe. Therefore, the algorithm reduces to an optimistic algorithm and the additive term becomes zero. Further discussion on this can be found in Corollary C.11.

The proof of Theorem 4.2 is provided in Appendix C. We outline the major steps of the proof as follows.

As discussed in Section 4.2, one pivotal component in StepMix is the construction of the candidate policies. As a result, in our proof, we first extend the good event related to $(\hat{P}_h^k - P_h)V_{h+1}^\star$ to $H+1$ good events related to $(\hat{P}_h^k - P_h)V_{h+1}^{\star,h_0}$, $h_0 = 0, 1, \ldots, H$. Since $\pi^{*,h_0}$ is a fixed anchor policy, $(\hat{P}_h^k - P_h)V_{h+1}^{\star,h_0}$ can be bounded for all $h_0$.

Next, we show that $\tilde{Q}^{k,h_0}$ and $\underset{\sim}{Q}^{k,h_0}$ are valid UCB and LCB of $Q^{k,h_0}$ and $Q^{\star,h_0}$ respectively, in the sense that

$\underset{\sim}{Q}^{k,h_0} \leq Q^{k,h_0} \leq Q^{\star,h_0} \leq \tilde{Q}^{k,h_0}$. Meanwhile, we show that $\tilde{Q}^{k,h_0}$ and $\underset{\sim}{Q}^{k,h_0}$ are sufficiently tight, as $\tilde{Q}^{k,h_0} - \underset{\sim}{Q}^{k,h_0}$ is bounded and will converge to zero sufficiently fast. Furthermore, with the properties of our constructed step mixture policy, we can obtain tight UCB and LCB for the step mixture policies as well.

Finally, we show that $\pi_k$ only stays at $\pi^b$ or the mixture policy for finite number of episodes. This is due to the fact that $V_1^{\star,h_0} \geq V_1^{\pi_b}$ for any $h_0 = 0, 1, \ldots, H$. Thus, with high probability, $\underset{\sim}{V}_1^{k,h_0} \geq \gamma$ when $k$ is sufficiently large. As a result, the agent will then select the optimistic policy $\bar{\pi}^k$ in most of the episodes. Thus, the regret of StepMix has the same leading term as that under the optimistic policy, which will then be bounded efficiently.

## 5. The EpsMix Algorithm

In this section, we briefly introduce another algorithm named EpsMix and defer the detailed design and analysis to Appendix D. Different from StepMix in Algorithm 1, EpsMix does not construct step mixture policies during the learning process. Rather, it adopts a randomization mechanism at the beginning of each episode, and designs episodic mixture policies (Wiering & Van Hasselt, 2008; Baram et al., 2021) defined as follows.

**Definition 5.1** (Episodic Mixture Policy)**.** Given two policies $\pi^1$ and $\pi^2$ with parameter $\rho \in (0, 1)$, the episodic mixture policy, denoted by $\rho\pi^1 \oplus (1-\rho)\pi^2$, randomly picks $\pi^1$ with probability $\rho$ and $\pi^2$ with probability $1-\rho$ at the beginning of an episode and plays it for the entire episode.

The EpsMix algorithm is presented in Algorithm 4 in Appendix D, and it proceeds as follows. Similar to StepMix, at the beginning of each episode $k$, it first constructs an optimistic policy, denoted as $\bar{\pi}^k$. It then evaluates the LCB of the expected total rewards under both $\bar{\pi}^k$ and $\pi^b$, denoted as $\underset{\sim}{V}_1^k$ and $\underset{\sim}{V}_1^{k,b}$ respectively. If $\underset{\sim}{V}_1^k$ is above the threshold $\gamma$, it indicates that the optimistic policy $\bar{\pi}^k$ satisfies the conservative constraint with high probability. The learner thus executes $\bar{\pi}^k$ in the following episode $k$. Otherwise, if $\underset{\sim}{V}_1^{k,b}$ is above the threshold while $\underset{\sim}{V}_1^k$ is not, it constructs an episodic mixture policy $\rho_k\bar{\pi}^k \oplus (1-\rho_k)\pi^b$ so that $\rho_k \underset{\sim}{V}_1^k + (1-\rho_k)\underset{\sim}{V}_1^{k,b} = \gamma$. It implies that the episodic policy satisfies the conservative constraint in expectation with high probability. If neither $\underset{\sim}{V}_1^k$ nor $\underset{\sim}{V}^{k,b}$ is above the threshold, EpsMix will resort to the baseline policy to collect more information.

Our theoretical analysis shows that EpsMix has the same performance guarantees as StepMix. At the same time, we note that EpsMix is less conservative than StepMix in the sense that, the expected return under a *selected* policy in an episode may be below the threshold when $\underset{\sim}{V}_1^k < \gamma$. However, when taking the randomness in the policy mixture procedure into consideration, we can still guarantee that the expected total return under an episodic mixture policy is above the threshold with probability at least $1-\delta$.

## 6. From Baseline Policy to Offline Dataset

Both EpsMix and StepMix critically depend on the baseline policy $\pi^b$ to achieve the desired conservative guarantee. In reality, however, a baseline policy that provably satisfies the conservative constraint may not always be explicitly given to the algorithm. Instead, the learning agent may have access to an offline dataset that is collected from the target environment by executing an unknown behavior policy $\mu$, and the goal is to design a conservative exploration algorithm that satisfies Equation (2) only using the offline dataset.

A natural approach to solve this problem is to first learn a baseline policy from the dataset, and then use it as an input to EpsMix or StepMix. The challenge, however, is that instead of having full confidence in the conservative guarantee of $\pi^b$, we must deal with the *safety uncertainty* of the learned baseline policy, that is introduced by using the offline dataset as well as the offline learning algorithm that produces the baseline policy. Fortunately, we prove that for StepMix, the uncertainty of learning a safe baseline policy from the offline dateset does not affect the conservative constraint violation or the regret order if the offline dataset is sufficiently large.

**Theorem 6.1.** *Let $\hat{\pi}$ be the output of the offline VI-LCB algorithm (Xie et al., 2021) (see Algorithm 5 in Appendix E) with $n = \tilde{\Theta}(\frac{H^5SA}{\bar{\kappa}^2})$[1] offline trajectories. If we replace the baseline policy $\pi^b$ used in Algorithm 1 by $\hat{\pi}$, then with probability at least $1-\delta$, StepMix can simultaneously (i) satisfy the conservative constraint in Equation (2), and (ii) achieve a total regret that is at most*

$$\tilde{O}\left(\sqrt{H^3SAK} + H^3S^2A + H^3SA\bar{\Delta}_0\left(\tfrac{1}{\bar{\kappa}^2} + \tfrac{S}{\bar{\kappa}}\right)\right),$$

*where $\bar{\kappa} = (V_1^\mu - \gamma)/2 > 0$ and $\bar{\Delta}_0 = V_1^\star - V_1^\mu + \bar{\kappa}$.*

A similar result for EpsMix can be established, and is given as Theorem E.7 in Appendix E. We see that $n$ scales inversely proportional to $\kappa^2$, suggesting that a good behavior policy would require small amount of data and vice versa. Besides, the additive term in the regret becomes larger compared with that in Theorem 4.2. In general, a large $n$ serves two purposes: First, it reduces the safety uncertainty due to offline learning, such that the impact on the safety constraint violation is negligible compared with that caused by the (online) StepMix policy. Second, it ensures that the regret bound is dominated by the number of online episodes $K$. We also note that although both Theorem 6.1 and The-

[1]We hide the logarithm factor for simplicity.

orem E.7 depend on using VI-LCB as the offline learning algorithm, the conclusion can be extended to general offline algorithms as long as they can produce an approximately safe policy from the pre-collected data with high probability.

# 7. Experimental Results

## 7.1. Performance Evaluation of StepMix and EpsMix

**Synthetic Environment.** We generate a synthetic environment to evaluate the proposed algorithms. We set the number of states $S$ to be 5, the number of actions $A$ for each state to be 5, and the episode length $H$ to be 3. The reward $r_h(s, a)$ for each state-action pair and each step is generated independently and uniformly at random from $[0, 1]$. We also generate the transition kernel $P_h(\cdot|s, a)$ from an $S$-dimensional simplex independently and uniformly at random. Such procedure guarantees that the synthetic environment is a proper tabular MDP.

**Baseline Policy.** We adopt the Boltzmann policy (Thrun, 1992) as the baseline policy in our algorithms. Under the Boltzmann policy, actions are taken randomly according to $\pi_h(a|s) = \frac{\exp\{\eta Q_h^\star(s,a)\}}{\sum_{a\in\mathcal{A}} \exp\{\eta Q_h^\star(s,a)\}}$, where a larger $\eta$ leads to a more deterministic policy and higher expected value.

**Results.** We first evaluate the proposed StepMix and EpsMix, and compare with BPI-UCBVI (Ménard et al., 2021). For each algorithm, we run 10 trials and plot the average expected return per episode.

In Figure 1, we track the expected return obtained in each episode with different baseline parameter $\eta$ and conservative constraint $\gamma$. We have the following observations. First, both StepMix and EpsMix converge to the optimal policy with no constraint violation in all settings. Between StepMix and EpsMix, the latter exhibits slightly faster convergence. They both tend to stay on the baseline policy when the information is not sufficient, implied by the constant expected return at the beginning of the learning process. When more information is collected, these two algorithms will deviate from the baseline policy and converge to the optimal policy. In contrast, BPI-UCBVI converges to the optimal policy as well, but violates the conservative constraints in earlier episodes. Besides, more stringent constraint $\gamma$ makes StepMix and EpsMix more conservative. Both algorithms experience delayed convergence when $\gamma$ increases. Meanwhile, a better baseline policy also leads to better learning performance throughout the learning process.

We report the performance of learning with an offline dataset in Figure 2. We use the baseline Boltzmann policy with $\eta = 10$ and $\eta = 15$ to collect the offline dataset. The numbers of offline trajectories are set to be $5000$ and $8000$, respectively. The conservative constraint $\gamma$ is set to be 2.2. Figure 2 shows that learning a baseline policy from the offline dataset and using it as an input to StepMix and EspMix does not affect their performances significantly. With more offline trajectories collected, the algorithms start from a better baseline and converge to the optimal policy faster.

## 7.2. Empirical Comparison with DOPE and OptPess-LP

In this subsection, we empirically compare the learning performances of StepMix, EpsMix, DOPE (Bura et al., 2022) and OptPess-LP (Liu et al., 2021).

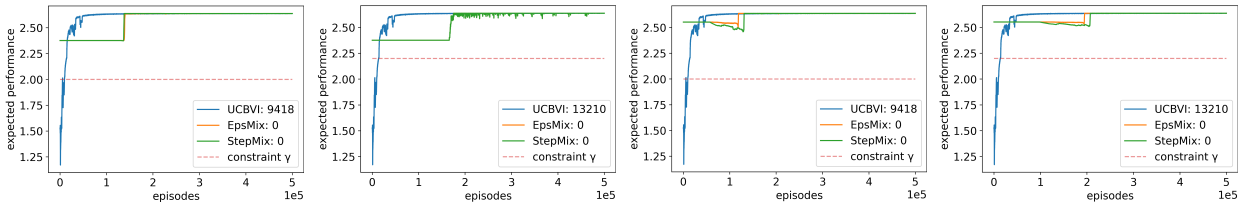
**Synthetic Homogeneous Environment.** In this experiment, we set $S$ to be 4, $A$ to be 2, and $H$ to be 3. In order to match the homogeneous environment assumption under DOPE and OptPess-LP, we set $P_h = P$ and $r_h = r$ for any $h \in [H]$, and randomly generate $P$ an $r$ as in Section 7.1. As DOPE and OptPess-LP are both developed to solve CMDP problems with general cost functions, to match the conservative constraint considered in this work, we set the corresponding cost function as $c(s, a) = 1 - r(s, a)$ and set the constraint as $\mathbb{E}_\pi[\sum_{h=1}^H c(s_h, a_h)] \leq H - \gamma$.

**Results.** We adopt the Boltzmann policy from Section 7.1 as the baseline policy and set $\eta$ to be 5. We run each algorithm for 10 trials and plot the average regrets in Figure 3(a) and the average expected return of each episode in Figure 3(b).

We observe that all four algorithms achieve the same performance at the beginning of the learning process, implying that they all adopt the baseline policy initially. After that, DOPE is the first algorithm to deviate from the baseline and explore other safe policies, followed by StepMix and EpsMix. Although DOPE starts the exploration earlier, it actually renders much higher regret than StepMix and EpsMix. This implies that the exploration under DOPE is not as efficient as the near-optimal exploration strategies adopted by StepMix and EpsMix. On the other hand, OptPess-LP stays on the baseline throughout the learning horizon, leading to a linearly increasing regret. This is because OptPess-LP does not explore sufficiently, and thus is unable to identify a safe exploration policy other than the baseline in this scenario. Similar phenomenon has been observed in Bura et al. (2022). Figure 3(b) also shows that the constraint violation is zero throughout the learning horizon under all four algorithms.

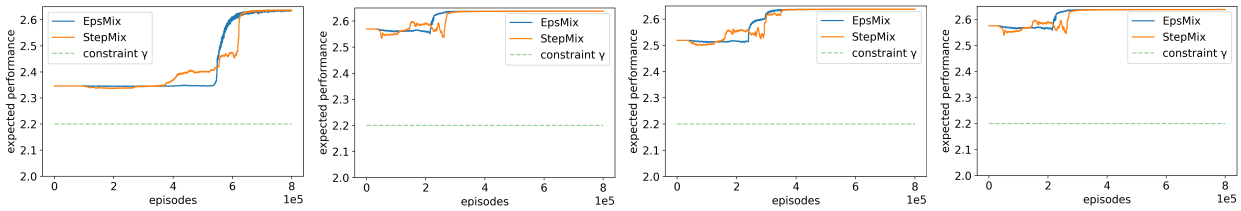
# 8. Conclusions

We investigated conservative exploration in episodic tabular MDPs. Different than the majority of existing literature, we considered a stringent episodic conservative constraint, which motivated us to incorporate mixture policies in conservative exploration. We proposed two model-based algorithms, one with step mixture policies and the other with episodic randomization. Both algorithms were proved to achieve near-optimal regret order as that under the constraint-free setting, while never violating the conserva-



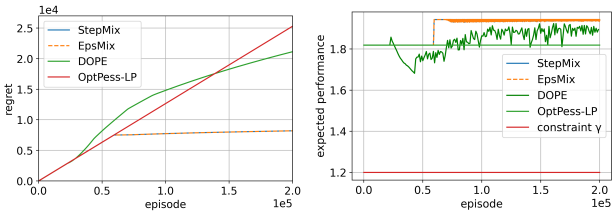

(a) $\eta = 5, \gamma = 2.0$ (b) $\eta = 5, \gamma = 2.2$ (c) $\eta = 10, \gamma = 2.0$ (d) $\eta = 10, \gamma = 2.2$

*Figure 1.* Average expected return of each episode under StepMix, EpsMix, and BPI-UCBVI with different constraint $\gamma$ and baseline parameter $\eta$. Numbers of violations are stated in the legend.




(a) $\eta = 10, n = 5000$ (b) $\eta = 10, n = 8000$ (c) $\eta = 15, n = 5000$ (d) $\eta = 15, n = 8000$

*Figure 2.* Average expected return of each episode under StepMix, EpsMix, and BPI-UCBVI with offline dataset.




(a) (b)

*Figure 3.* Performance comparison between StepMix, EpsMix, DOPE, and OptPess-LP. (a) Average regret. (b) Average expected return per episode.

tive constraint in the learning process. We also investigated a practical case where the baseline policy is not explicitly given to the algorithm, but must be learned from an offline dataset. We showed that as long as the dataset is sufficiently large, the offline learning step does not affect the conservative constraint or the regret of our proposed algorithms. Experimental results in a synthetic environment corroborated the theoretical analysis and shed some interesting light on the behavior of our algorithms.

## Acknowledgements

The work of DL, RH and JY was supported in part by the US National Science Foundation (NSF) under awards 2030026, 2003131, and 1956276. The work of CS was supported in part by the US NSF under awards 2143559, 2029978, 2002902, and 2132700.

## References

Altman, E. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

Amani, S., Alizadeh, M., and Thrampoulidis, C. Linear stochastic bandits under safety constraints. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Amani, S., Thrampoulidis, C., and Yang, L. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 243–253. PMLR, 2021.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Baram, N., Tennenholtz, G., and Mannor, S. Maximum entropy reinforcement learning with mixture policies. *arXiv preprint arXiv:2103.10176*, 2021.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(65):3207–3260, 2013.

Bura, A., Hasanzadezonuzy, A., Kalathil, D., Shakkottai, S., and Chamberland, J.-F. DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Dann, C., Lattimore, T., and Brunskill, E. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

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

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

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

Domingues, O. D., Ménard, P., Pirotta, M., Kaufmann, E., and Valko, M. Kernel-based reinforcement learning: A finite-time analysis. In *International Conference on Machine Learning*, pp. 2783–2792. PMLR, 2021b.

Du, Y., Wang, S., and Huang, L. A one-size-fits-all solution to conservative bandit problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7254–7261, 2021.

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

Garcelon, E., Ghavamzadeh, M., Lazaric, A., and Pirotta, M. Conservative exploration in reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1431–1441. PMLR, 2020a.

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

Garcelon, E., Ghavamzadeh, M., Lazaric, A., and Pirotta, M. Conservative exploration in reinforcement learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 1431–1441, 26–28 Aug 2020c.

Garcıa, J. and Fernández, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Ghosh, A., Zhou, X., and Shroff, N. Provably efficient model-free constrained RL with linear function approximation. *arXiv preprint arXiv:2206.11889*, 2022.

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

Jonsson, A., Kaufmann, E., Ménard, P., Darwiche Domingues, O., Leurent, E., and Valko, M. Planning in Markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 33:1253–1263, 2020.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Kalagarla, K. C., Jain, R., and Nuzzo, P. A sample-efficient algorithm for episodic finite-horizon MDP with constraints. *arXiv preprint arXiv:2009.11348*, 2020.

Kazerouni, A., Ghavamzadeh, M., Abbasi Yadkori, Y., and Van Roy, B. Conservative contextual linear bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

Khezeli, K. and Bitar, E. Safe linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10202–10209, 2020.

Laroche, R., Trichelair, P., and Combes, R. T. D. Safe policy improvement with baseline bootstrapping. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., and Tian, C. Learning policies with zero or bounded constraint violation for constrained MDPs. *Advances in Neural Information Processing Systems*, 34, 2021.

Luo, H., Wei, C.-Y., and Lee, C.-W. Policy optimization in adversarial MDPs: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34:22931–22942, 2021.

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

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

Petrik, M., Ghavamzadeh, M., and Chow, Y. Safe policy improvement by minimizing robust baseline regret. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2306–2314, 2016.

Pirotta, M., Restelli, M., Pecorino, A., and Calandriello, D. Safe policy iteration. In *International Conference on Machine Learning*, pp. 307–315. PMLR, 2013.

Qiu, S., Wei, X., Yang, Z., Ye, J., and Wang, Z. Upper confidence primal-dual optimization: Stochastically constrained Markov decision processes with adversarial losses and unknown transitions. *arXiv preprint arXiv:2003.00660*, 2020.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

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

Simão, T. D. and Spaan, M. T. J. Safe policy improvement with baseline bootstrapping in factored environments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:4967–4974, Jul. 2019.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 814–823, Lille, France, 07–09 Jul 2015. PMLR.

Tamar, A., Di Castro, D., and Mannor, S. Policy gradients with variance related risk criteria. In *Proceedings of the 29th International Conference on Machine Learning*, ICML'12, pp. 1651–1658, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2380–2388, Lille, France, 07–09 Jul 2015a.

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. High confidence off-policy evaluation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 3000–3006. AAAI Press, 2015b. ISBN 0262511290.

Thrun, S. B. Efficient exploration in reinforcement learning. Technical report, Carnegie Mellon University, USA, 1992.

Turchetta, M., Kolobov, A., Shah, S., Krause, A., and Agarwal, A. Safe reinforcement learning via curriculum induction. *arXiv preprint arXiv:2006.12136*, 2020.

Wei, H., Liu, X., and Ying, L. Triple-Q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3274–3307. PMLR, 2022.

Wiering, M. A. and Van Hasselt, H. Ensemble algorithms in reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(4):930–936, 2008.

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

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

Xie, T., Bhardwaj, M., Jiang, N., and Cheng, C.-A. Armor: A model-based framework for improving arbitrary baseline policies with offline data. *arXiv preprint arXiv:2211.04538*, 2022.

Xu, T., Liang, Y., and Lan, G. A primal approach to constrained policy optimization: Global optimality and finite-time analysis. *arXiv preprint arXiv:2011.05869*, 2020.

Yang, T.-Y., Rosca, J., Narasimhan, K., and Ramadge, P. J. Accelerating safe reinforcement learning with constraint-mismatched baseline policies. In *International Conference on Machine Learning*, pp. 11795–11807. PMLR, 2021.

Yang, Y., Wu, T., Zhong, H., Garcelon, E., Pirotta, M., Lazaric, A., Wang, L., and Du, S. S. A reduction-based framework for conservative bandits and reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=AcrlgZ9BKed.

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

Zheng, L. and Ratliff, L. J. Constrained upper confidence reinforcement learning. *arXiv preprint arXiv:2001.09377*, 2020.

Zhong, H., Yang, Z., Wang, Z., and Szepesvári, C. Optimistic policy optimization is provably efficient in non-stationary MDPs. *arXiv preprint arXiv:2110.08984*, 2021.

# A. Related Works

**Constrained RL with Baseline Policies.** Conservative exploration studied in this paper can be viewed as a specific case of the Constrained Markov Decision Process (CMDP) (Altman, 1999), which has been investigated in both offline and online settings. In the *offline* setting, a given baseline policy produces a set of trajectories for the agent to learn a policy that is guaranteed to perform at least as good as the baseline with high probability without actually interacting with the MDP (Bottou et al., 2013; Thomas et al., 2015b;a; Swaminathan & Joachims, 2015; Petrik et al., 2016; Laroche et al., 2019; Simão & Spaan, 2019). It can also be extended to the *semi-batch* setting (Pirotta et al., 2013). In the *online* setting, the agent has to trade off exploration and exploitation while interacting with the MDP. Several algorithms have been proposed in the literature (Garcelon et al., 2020c; Yang et al., 2022). Garcelon et al. (2020c) introduce a Conservative Upper-Confidence Bound for Reinforcement Learning (CUCRL2) algorithm for both finite horizon and average reward problems with $O(\sqrt{T})$ regret. Yang et al. (2022) propose a reduction-based framework for conservative bandits and RL, which translates a minimax lower bound of the non-conservative setting to a valid lower bound for the conservative case. They also propose a Budget-Exploration algorithm and show that its regret scales in $\tilde{O}\left(\sqrt{H^3SAK} + \frac{H^3SA\Delta_0}{\kappa(\kappa+\Delta_0)}\right)$ for tabular MDPs, where $\Delta_0$ is the suboptimality gap of the baseline policy, and $\kappa$ is the tolerable performance loss from the baseline. However, all these works assume *cumulative* conservative constraint.

**Other Forms of Constraints.** Beside the constraint imposed by a baseline policy, which is generally "aligned" with the learning goal, CMDP also studies the case where the algorithm must satisfy a set of constraints that potentially are not aligned with the reward. In general, both cumulative cost constraints (Efroni et al., 2020; Turchetta et al., 2020; Zheng & Ratliff, 2020; Qiu et al., 2020; Ding et al., 2020; Kalagarla et al., 2020; Liu et al., 2021; Wei et al., 2022; Ghosh et al., 2022) and episodic cost constraints (Liu et al., 2021; Bura et al., 2022; Huang et al., 2022) have been investigated. Assuming a known safe baseline policy that satisfies the corresponding constraints, OptPess-LP (Liu et al., 2021) is shown to achieve a regret of $\tilde{O}(\frac{1}{\kappa}\sqrt{H^6S^3AK})$ without any constraint violation with high probability, while DOPE (Bura et al., 2022) improves the regret to $\tilde{O}(\frac{1}{\kappa}\sqrt{H^6S^2AK})$, where $\kappa$ denotes the Slater parameter. We note that both algorithms do not achieve the optimal regret in the unconstrained counterpart, due to the adopted linear programming-based approaches. Beyond tabular setting, CMDP has also been discussed in linear (Ding et al., 2021; Ghosh et al., 2022; Amani et al., 2021; Yang et al., 2022) or low-rank models (Huang et al., 2022). Other formulations different from conservative exploration or CMDP, such as minimizing the variance of expected return (Tamar et al., 2012) or generally, maximizing some utility function of state-action pairs (Ding et al., 2021), have also been investigated. Lastly, Yang et al. (2021) study constrained reinforcement learning with a baseline policy that may not satisfy the given set of constraints.

**Safe Bandits.** Bandits problem is a standard RL problem where it interacts with a stationary environment, which reduces the difficulties of learning. Several constraints are considered in the bandits setting. The first is that the cumulative expected reward of an agent should exceed a certain threshold. This setting is originally studied in Wu et al. (2016), which adopts an UCB type of exploration and checks whether the policy satisfies the conservative constraint. Kazerouni et al. (2017); Garcelon et al. (2020b); Pacchiano et al. (2021) then extend the conservative setting to contextual linear bandits. The second constraint is much stronger, as it requires that each arm played by the learning agent be safe given the baseline or the threshold. Amani et al. (2019) and Khezeli & Bitar (2020) both use an LCB type of algorithm to ensure the arms selected by the algorithms are safe under linear bandits setting. Du et al. (2021) consider conservative exploration with a sample-path constraint on the actual observed rewards rather than in expectation.

**Policy Optimization.** This is another research direction in RL that utilizes baseline policies (Schulman et al., 2015). However, the focus and assumptions of these papers are very different from this work. For example, Zhong et al. (2021) and Luo et al. (2021) focus on the non-stationary and adversary environments, respectively. While policy optimization can achieve sublinear regret under certain MDP models (Shani et al., 2020), it usually lacks performance guarantees during the learning process, which is in stark contrast to our results.

**Other LCB Techniques.** We highlight the differences between the LCBs used in our online algorithms StepMix and EpsMix and two LCB techniques used in Xie et al. (2021; 2022). First and foremost, Xie et al. (2021; 2022) study offline RL problems. They both impose a coverage assumption on the behavior policy in order to bound the estimation error of the pessimistic policy constructed from the offline dataset. On the other hand, our LCB has no such coverage assumption for the behavior policy, since the LCB constructed in our work is a lower bound of the value function under the online exploration policy $\pi^k$. Second, the reference value functions in this work are different than those in Xie et al. (2021), which constructs a reference function based on an LCB algorithm so that it can derive the pessimistic policy and utilize Bernstein's inequality. In our work, however, the reference functions are the true value functions of candidate policies, which are chosen from a

series of policies mixed by the baseline policy and optimistic policy produced from a Bernstein-style UCB algorithm. Last but not the least, our LCB expression is more computationally efficient than that in Xie et al. (2022).

## B. Notations

We list the notations of common quantities as follows.

| Notation | Meaning | Definition |
|---|---|---|
| $r_h(s,a)$ | reward | - |
| $P_h(s'\|s,a)$ | transition probability | - |
| $n_h^k(s,a)$ | visitation count | $\sum_{\tau=1}^{k-1} \mathbb{1}\{s_h^\tau = s, a_h^\tau = a\}$ |
| $n_h^k(s,a,s')$ | visitation-transition count | $\sum_{\tau=1}^{k-1} \mathbb{1}\{s_h^\tau = s, a_h^\tau = a, s_{h+1}^\tau = s'\}$ |
| $\hat{P}_h^k(s'\|s,a)$ | empirical estimate of transition probability | $\frac{n_h^k(s,a,s')}{n_h^k(s,a)}$ if $n_h^k(s,a) \geq 1$; $\frac{1}{S}$, otherwise |
| $d_h^\pi(s,a)$ | occupancy measure under policy $\pi$ | $\mathbb{E}_\pi[\mathbb{1}\{s_h = s, a_h = a\}]$ |
| $d_h^k(s,a)$ | occupancy measure under policy $\pi^k$ | $d_h^{\pi^k}(s,a)$ |
| $\bar{n}_h^k(s,a)$ | expected visitation count | $\sum_{\tau=1}^{k-1} d_h^\tau(s,a)$ |
| $Q_h^\pi(s,a)$ | true Q function | $\mathbb{E}_\pi[\sum_{i=h}^H r_i(s_i,a_i)\|s_h = s, a_h = a]$ |
| $V_h^\pi(s)$ | true V function | $\mathbb{E}_\pi[\sum_{i=h}^H r_i(s_i,a_i)\|s_h = s]$ |
| $\pi^\star$ | optimal policy | $\arg\max_\pi V_1^\pi(s_1)$ |
| $\pi^b$ | baseline policy | - |
| $\pi^{\star,h_0}$ | step-wise optimal policy | $\{\pi_1^b, \cdots, \pi_{h_0}^b, \pi_{h_0+1}^\star, \cdots, \pi_H^\star\}$ |
| $\bar{\pi}^k$ | global optimistic policy | constructed from Algorithm 2 |
| $\pi^{k,h_0}$ | step-wise optimistic policy | $\{\pi_1^b, \cdots, \pi_{h_0}^b, \bar{\pi}_{h_0+1}^k, \cdots, \bar{\pi}_H^k\}$ |
| $Q^{k,h_0}, V^{k,h_0}, Q^{\star,h_0}, V^{\star,h_0}$ | corresponding true value functions | $Q^{\pi^{k,h_0}}, V^{\pi^{k,h_0}}, Q^{\pi^{\star,h_0}}, V^{\pi^{\star,h_0}}$ |
| $\tilde{Q}^{k,h_0}, \tilde{V}^{k,h_0}$ | Upper Confidence Bounds | defined in Equations (8) and (10) |
| $\underset{\sim}{Q}^{k,h_0}, \underset{\sim}{V}^{k,h_0}$ | Lower Confidence Bounds | defined in Equations (9) and (11) |
| $G^{k,h_0}$ | G function | defined in Equation (24) |
| $\beta(n,\delta)$ | logarithm term involved in $\mathcal{E}$ | $\log(SAH/\delta) + S\log(8e(n+1))$ |
| $\beta^{\text{cnt}}(\delta)$ | logarithm term involved in $\mathcal{E}^{\text{cnt}}$ | $\log(SAH/\delta)$ |
| $\beta^\star(n,\delta)$ | logarithm term involved in $\mathcal{E}^\star$ | $\log(SAH/\delta) + \log(8e(n+1))$ |

We also adopt the following $\min, \max$ notations:

$$a \wedge b = \min(a,b), a \vee b = \max(a,b).$$

For any given policy $\pi$ and $Q$ function, we denote

$$\pi_h Q_h(s) = \langle \pi_h(\cdot|s), Q_h(s,\cdot)\rangle = \sum_{a\in\mathcal{A}} \pi_h(a|s) Q_h(s,a).$$

For any given transition kernel $P_h$ and value function $V_{h+1}$, we define the variance of $P_h V_{h+1}(s,a)$ as follows.

$$\mathrm{Var}_{P_h}(V_{h+1})(s,a) = \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[(V_{h+1}(s') - \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}(s')])^2].$$

At last, we introduce three types of good events and their notations that will be intensively used in the following proofs.

The first type of good events characterizes the connection between the true visitation counts and the expected visitation counts:

$$\mathcal{E}^{\text{cnt}}(\delta) \triangleq \left\{\forall k \in [K], \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : n_h^k(s,a) \geq \frac{1}{2}\bar{n}_h^k(s,a) - \beta^{\text{cnt}}(\delta)\right\},$$

where $\beta^{\text{cnt}} = \log(SAH/\delta)$, $n_h^k(s,a)$ denotes the number of visitations of state-action pair $(s,a)$ and $\bar{n}_h^k(s,a)$ denotes the expected visitation count.

The second type of good events, defined as follows, upper bounds the KL divergence between the estimated transition distribution and the true transition distribution.

$$\mathcal{E}(\delta) \triangleq \left\{\forall k \in [K], \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \text{KL}(\hat{P}_h^k(\cdot|s,a), P_h(\cdot|s,a)) \leq \frac{\beta(n_h^k(s,a),\delta)}{n_h^k(s,a)}\right\},$$

where $\beta(n,\delta) = \log(SAH/\delta) + S\log(8e(n+1))$.

The third type of good events provides a Bernstein-style concentration guarantee, defined as follows.

$$\mathcal{E}^\star(V,\delta) \triangleq \Bigg\{\forall k \in [K], \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \in [H] \cup \{0\} : \left|(\hat{P}_h^k - P_h)V_{h+1}(s,a)\right| \leq \min\left\{H, \sqrt{2\text{Var}_{P_h}(V_{h+1})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta)}{n_h^k(s,a)}} + 3H\frac{\beta^\star(n_h^k(s,a),\delta)}{n_h^k(s,a)}\right\}\Bigg\},$$

where $\beta^\star(n,\delta) = \log(SAH/\delta) + \log(8e(n+1))$ and $V$ is a value function independent with $\hat{P}_h^k$ and bounded by $H$. Later, $V$ will be chosen separately in StepMix or EpsMix.

## C. Algorithm Design and Analysis of StepMix

We first recall the StepMix algorithm (Algorithm 2) and provide the PolicyEva subroutine in Algorithm 3.

Based on the construction of $\pi^{k,h_0}$ in Algorithm 2, we define the following Q-value functions and value functions.

$$\begin{aligned}\tilde{Q}_h^{k,h_0}(s,a) \triangleq \min\Bigg(&H, r_h(s,a) + 3\sqrt{\text{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 14H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\ &+ \frac{1}{H}\hat{P}_h^k(\tilde{V}_{h+1}^{k,h_0} - \underset{\sim}{V}_{h+1}^{k,h_0})(s,a) + \hat{P}_h^k\tilde{V}_{h+1}^{k,h}(s,a)\Bigg)\end{aligned} \tag{8}$$

$$\begin{aligned}\underset{\sim}{Q}_h^{k,h_0}(s,a) \triangleq \max\Bigg(&0, r_h(s,a) - 3\sqrt{\text{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} - 22H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\ &- \frac{2}{H}\hat{P}_h^k(\tilde{V}_{h+1}^{k,h_0} - \underset{\sim}{V}_{h+1}^{k,h_0})(s,a) + \hat{P}_h^k\underset{\sim}{V}_{h+1}^{k,h_0}(s,a)\Bigg)\end{aligned} \tag{9}$$

$$\tilde{V}_h^{k,h_0}(s) \triangleq \langle \pi_h^{k,h_0}(\cdot|s), \tilde{Q}_h^{k,h_0}(s,\cdot)\rangle \tag{10}$$

$$\underset{\sim}{V}_h^{k,h_0}(s) \triangleq \langle \pi_h^{k,h_0}(\cdot|s), \underset{\sim}{Q}_h^{k,h_0}(s,\cdot)\rangle. \tag{11}$$

We point out that, since the definitions of Equations (8) and (9) are the same as Equation (5) used in Algorithm 2, $\underset{\sim}{V}^{k,h_0}$ defined in Equation (11) is consistent with the same quantity used in Algorithm 2. Moreover, when $h > h_0$, we have $\pi_h^{k,h_0} = \bar{\pi}_h^k$, which implies that $\tilde{Q}_h^{k,h_0} = \tilde{Q}_h^k$ and $\pi_h^{k,h_0}(s) = \arg\max_a \underset{\sim}{Q}_h^{k,h_0}(s,a)$.

Before proceeding to the formal proof, we outline its major steps as follows.

**Step one:** In Appendix C.1, we verify that the good events happen with high probability and introduce the linearity of the ocupancy measure and the value function of the step-mix policies.

**Algorithm 2** The StepMix Algorithm

**Input:** $\pi^b$, $\gamma$, $\beta$, $\beta^\star$, $\mathcal{D}_0 = \emptyset$.
**for** $k$ = 1 to $K$ **do**
  Update model estimate $\hat{P}$ according to Equation (4).
  *# Optimistic policy identification*
  $\tilde{V}^k_{H+1}(s) = \underset{\sim}{V}^k_{H+1}(s) = 0, \forall s \in \mathcal{S}$.
  **for** $h$ = $H$ to 1 **do**
    Update $\tilde{Q}^k_h(s,a)$, $\underset{\sim}{Q}^k_h(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$ according to Equation (5).
    $\bar{\pi}^k_h(s) \leftarrow \arg\max_a \tilde{Q}^k_h(s,a)$, $\tilde{V}^k_h(s) \leftarrow \tilde{Q}^k_h(s, \bar{\pi}^k_h(s))$, $\underset{\sim}{V}^k_h(s) \leftarrow \underset{\sim}{Q}^k_h(s, \bar{\pi}^k_h(s)), \forall s \in \mathcal{S}$.
  **end for**
  *# Candidate policy construction and evaluation*
  **for** $h_0$ = 0 to $H$ **do**
    $\pi^{k,h_0} = \{\pi^b_1, \pi^b_2, \cdots, \pi^b_{h_0}, \bar{\pi}^k_{h_0+1}, \cdots, \bar{\pi}^k_{H-1}, \bar{\pi}^k_H\}$.
    $\underset{\sim}{V}^{k,h_0} = \text{PolicyEva}(\hat{P}^k, \pi^{k,h_0})$.
  **end for**
  *# Safe exploration policy selection*
  **if** $\{h \mid \underset{\sim}{V}^{k,h}_1 \geq \gamma, h = 0, 1, \ldots, H\} = \emptyset$ **then**
    $\pi^k = \pi^b$.
  **else**
    $h^k = \min\{h \mid \underset{\sim}{V}^{k,h}_1 \geq \gamma, h = 0, 1, \ldots, H\}$.
    **if** $h^k = 0$ **then**
      $\pi^k = \bar{\pi}^k$.
    **else**
      Set $\pi^k$ according to Equation (7).
    **end if**
  **end if**
  Execute $\pi^k$ and collect $\{(s^k_h, a^k_h, s^k_{h+1})\}_{h=1}^H$.
  $\mathcal{D}_n \leftarrow \mathcal{D}_{n-1} \cup \{(s^k_h, a^k_h, s^k_{h+1})\}_{h=1}^H$.
**end for**

**Algorithm 3** PolicyEva Subroutine

**Input:** $\hat{P}^k$, $\pi$
**Initialization:** Set $\tilde{V}^k_{H+1}(s)$ and $\underset{\sim}{V}^k_{H+1}(s)$ to be 0 for any $s \in \mathcal{S}$.
**for** $h$ = $H$ to 1 **do**
  Update $\tilde{Q}^k_h(s,a)$, $\underset{\sim}{Q}^k_h(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$:

$$\tilde{Q}^k_h(s,a) \triangleq \min\left(H, r_h(s,a) + 3\sqrt{\mathrm{Var}_{\hat{P}^k_h}(\tilde{V}^k_{h+1})(s,a)\frac{\beta^\star}{n^k_h(s,a)}} + 14H^2\frac{\beta}{n^k_h(s,a)} + \frac{1}{H}\hat{P}^k_h(\tilde{V}^k_{h+1} - \underset{\sim}{V}^k_{h+1})(s,a) + \hat{P}^k_h\tilde{V}^k_{h+1}(s,a)\right)$$

$$\underset{\sim}{Q}^k_h(s,a) \triangleq \max\left(0, r_h(s,a) - 3\sqrt{\mathrm{Var}_{\hat{P}^k_h}(\tilde{V}^k_{h+1})(s,a)\frac{\beta^\star}{n^k_h(s,a)}} - 22H^2\frac{\beta}{n^k_h(s,a)} - \frac{2}{H}\hat{P}^k_h(\tilde{V}^k_{h+1} - \underset{\sim}{V}^k_{h+1})(s,a) + \hat{P}^k_h\underset{\sim}{V}^k_{h+1}(s,a)\right)$$

  $\tilde{V}^k_h(s) \leftarrow \langle \tilde{Q}^k_h(s,\cdot), \pi_h(\cdot|s)\rangle$, $\underset{\sim}{V}^k_h(s) \leftarrow \langle \underset{\sim}{Q}^k_h(s,\cdot), \pi_h(\cdot|s)\rangle, \forall s \in \mathcal{S}$.
**end for**
**Output:** $\underset{\sim}{V}_1$

**Step two:** In Appendix C.2, we prove that $\tilde{Q}^{k,h_0}_h$ and $\underset{\sim}{Q}^{k,h_0}_h$ are valid UCB and LCB of the true Q-value functions of both the step-wise optimal policies $\pi^{\star,h_0}$ and the step-wise optimistic policies $\pi^{k,h_0}$, respectively, and provide a bound for the gap between $\tilde{Q}^{k,h_0}_h$ and $\underset{\sim}{Q}^{k,h_0}_h$.

**Step three:** In Appendix C.3, we leverage the bound for the gap between $\tilde{Q}^{k,h_0}_h$ and $\underset{\sim}{Q}^{k,h_0}_h$ to show a sublinear "weak" regret of the online policies $\pi^k$, where the regret is defined in terms of the performance difference $V^{\pi^b} - V^{\pi^k}$. Hence,

we can prove that there are only finite episodes in which the executed policy is not equal to the optimistic policy (finite non-optimistic policy lemma; cf. Lemma C.9).

**Step four:** In Appendix C.4, based on the bound of the gap between $\tilde{Q}_h^{k,h_0}$ and $\underset{\sim}{Q}_h^{k,h_0}$ and the finite non-optimistic policy lemma, we prove the regret stated in Theorem 4.2.

### C.1. Step One: Good Events and Basic Properties of Step Mixture Policies

We first prove the following lemma which shows that the good events defined in Appendix B occur with high probability.

**Lemma C.1** (Good Events)**.** *Let $\mathcal{E}, \mathcal{E}^{cnt}$, and $\mathcal{E}^\star$ be the events defined in Appendix B. Then, under Algorithm 1, with probability at least $1-\delta$, the following good events occur simultaneously:*

$$\mathcal{E}\left(\frac{\delta}{3}\right), \mathcal{E}^{cnt}\left(\frac{\delta}{3}\right), \mathcal{E}^\star\left(V^{\star,h_0}, \frac{\delta}{3(H+1)}\right), \forall h_0 \in [H] \cup \{0\}.$$

*Proof.* From Theorem F.7, Theorem F.8, and Theorem F.9, we have $\mathcal{E}(\frac{\delta}{3})$ , $\mathcal{E}^{\text{cnt}}(\frac{\delta}{3})$, and $\cap_{h_0 \in [H]\cup\{0\}} \mathcal{E}^\star(V^{\star,h_0}, \frac{\delta}{3(H+1)})$ occur with probability at least $1-\delta/3$, respectively. Then, by taking a union bound, all those good events occur simultaneously with probability at least $1-\delta$. □

Then, we provide several useful lemmas that capture the favorable properties of the step mixture policies $\pi^{k,h_0}$ and step-wise optimal policies $\pi^{*,h_0}$.

The following lemma establishes the optimality of policy $\pi^{\star,h_0}$.

**Lemma C.2** (Optimality of Step-wise Optimal Policies)**.** *Define $\Pi_{h_0} := \{\pi | \pi_h = \pi_h^b, \forall h \le h_0\}$. Then, for any $\pi \in \Pi_{h_0}$, we must have*

$$Q_h^{\star,h_0}(s,a) \ge Q_h^\pi(s,a),$$
$$V_h^{\star,h_0}(s) \ge V_h^\pi(s).$$

*Proof.* Using the performance difference lemma (Kakade & Langford, 2002), for any $\pi \in \Pi_{h_0}$ we have

$$V_h^\pi(s) - V_h^{\star,h_0}(s) = \sum_{m=h}^{H} \mathbb{E}_\pi[Q_m^{\star,h_0}(s_m,a_m) - V_m^{\star,h_0}(s_m) | s_h = s].$$

For the case where $h > h_0$, $\pi_h^{\star,h_0} = \pi_h^\star$, and thus $Q_h^{\star,h_0}(s,a) = Q_h^\star(s,a), \forall h \ge h_0, \forall s, a$. In addition, $\pi^\star(s) = \arg\max_a Q^\star(s,a)$. Thus, $\mathbb{E}_\pi[Q_h^{\star,h_0}(s_h,a_h) - V^{\star,h_0}(s_h)] \le 0, \forall h \ge h_0$.

For the case where $h \le h_0$, we have $\pi_h = \pi_h^{\star,h_0} = \pi_h^b$, $\mathbb{E}_\pi[Q_h^{\star,h_0}(s_h,a_h) - V^{\star,h_0}(s_h)] = 0, \forall h < h_0$.

Combining the results for both cases, we have

$$V_h^\pi(s) - V_h^{\star,h_0}(s) = \sum_{m=h}^{H} \mathbb{E}_\pi[Q_m^{\star,h_0}(s_m,a_m) - V_m^{\star,h_0}(s_m) | s_h = s] \le 0.$$

Following the same argument, we can prove that $Q_h^{\star,h_0}(s,a) \ge Q_h^\pi(s,a)$. □

The next lemma characterizes the property of the step mixture policies obtained by mixing two policies that are one-step different.

**Lemma C.3.** *If $\pi = \rho\pi^1 + \rho\pi^2$, where $\pi^1$ and $\pi^2$ differ only at step $h_0$, and let $d_h^1(s,a)$ and $d_h^2(s,a)$ be the occupancy measure of $\pi^1$ and $\pi^2$. Then we have*

$$d_h^\pi(s,a) = \rho d_h^1(s,a) + (1-\rho) d_h^2(s,a),$$

*where $d_h^\pi(s,a) = \mathbb{E}_\pi[\mathbb{1}\{s_h = s, a_h = a\}]$ is the occupancy measure under policy $\pi$.*

*Proof.* Based on the definition of $\pi$, we consider the following possible cases.

When $h < h_0$, we have $\pi_h^1 = \pi_h^2 = \pi_h$. Thus the corresponding occupancy measures should also be the same, i.e.,

$$d_h^1(s,a) = d_h^2(s,a) = d_h^\pi(s,a) = \rho d_h^1(s,a) + (1-\rho) d_h^2(s,a).$$

When $h = h_0$, we have $\pi_{h_0} = \rho\pi_{h_0}^1 + (1-\rho)\pi_{h_0}^2$. Using the fact that $d_{h_0-1}^1(s,a) = d_{h_0-1}^2(s,a) = d_{h_0-1}^\pi(s,a)$, we have

$$\begin{aligned}
d_{h_0}^\pi(s,a) =& \sum_{s',a'} \pi_{h_0}(a|s) P_{h_0-1}(s|s',a') d_{h_0-1}^\pi(s',a') \\
=& \sum_{s',a'} (\rho\pi_{h_0}^1(a|s) + (1-\rho)\pi_{h_0}^2(a|s)) P_{h_0-1}(s|s',a') d_{h_0-1}^\pi(s',a') \\
=& \rho \sum_{s',a'} \pi_{h_0}^1(a|s) P_{h_0-1}(s|s',a') d_{h_0-1}^1(s',a') + (1-\rho) \sum_{s',a'} \pi_{h_0}^2(a|s) P_{h_0-1}(s|s',a') d_{h_0-1}^2(s',a') \\
=& \rho d_{h_0}^1(s,a) + (1-\rho) d_{h_0}^2(s,a).
\end{aligned}$$

When $h > h_0$, we again have $\pi_h^1 = \pi_h^2 = \pi_h$. We then prove the equality through induction. Assume $d_{h-1}^\pi(s,a) = \rho d_{h-1}^1(s,a) + (1-\rho) d_{h-1}^2(s,a), \forall h-1 \geq h_0$, which holds when $h-1 = h_0$ based on the analysis above. Then,

$$\begin{aligned}
d_h(s,a) =& \sum_{s',a'} \pi_h(a|s) P_{h-1}(s|s',a') d_{h-1}^\pi(s',a') \\
=& \sum_{s',a'} \pi_h(a|s) P_{h-1}(s|s',a') (\rho d_{h-1}^1(s,a) + (1-\rho) d_{h-1}^2(s,a)) \\
=& \rho \sum_{s',a'} \pi_h^1(a|s) P_{h-1}(s|s',a') d_{h-1}^1(s,a) + (1-\rho) \sum_{s',a'} \pi_h^2(a|s) P_{h-1}(s|s',a') d_{h-1}^2(s,a) \\
=& \rho d_h^1(s,a) + (1-\rho) d_h^2(s,a),
\end{aligned}$$

which completes the proof. □

The above lemma shows that the occupancy measure of a step mixture policy obtained by mixing two policies that are one-step different is a linear combination of the occupancy measures of the corresponding policies. Such linearity also holds for the corresponding value functions, as shown in the following proposition.

**Proposition C.4.** *With the same condition as in Lemma C.3, the following equality holds:*

$$V_1^\pi = \rho V_1^{\pi^1} + (1-\rho) V_1^{\pi^2}.$$

*Proof.* By the definition of $V_1^\pi$ and $d_h^\pi(s,a)$, we have $V_1^\pi = \sum_{h=1}^H \sum_{s,a} d_h^\pi(s,a) r_h(s,a)$. Hence,

$$\begin{aligned}
V_1^\pi =& \sum_{h=1}^H \sum_{s,a} d_h^\pi(s,a) r_h(s,a) \\
=& \sum_{h=1}^H \sum_{s,a} \left(\rho d_h^1(s,a) + (1-\rho) d_h^2(s,a)\right) r_h(s,a) \\
=& \rho V_1^{\pi^1} + (1-\rho) V_1^{\pi^2}.
\end{aligned}$$

□

### C.2. Step Two: Confidence Bounds

In this step, we validate the UCBs and LCBs constructed in Equations (8) to (11), and provide bounds for the estimation error induced by the pairs of UCBs and LCBs.

Recall that the upper confidence bounds of value functions of each policy $\pi^{k,h_0}$ are

$$\begin{cases} \tilde{Q}_h^{k,h_0}(s,a) \triangleq \min\Bigg(H, r_h(s,a) + 3\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 14H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\ \qquad + \frac{1}{H}\hat{P}_h^k(\tilde{V}_{h+1}^{k,h_0} - \underset{\sim}{V}_{h+1}^{k,h_0})(s,a) + \hat{P}_h^k\tilde{V}_{h+1}^{k,h_0}(s,a)\Bigg), \\ \tilde{V}_h^{k,h_0}(s) \triangleq \langle \pi_h^{k,h_0}(\cdot|s), \tilde{Q}_h^{k,h_0}(s,\cdot)\rangle, \end{cases} \tag{12}$$

where $\delta' = \frac{\delta}{3(H+1)}$ and

$$\pi_h^{k,h_0}(s) = \begin{cases} \pi_h^b(s), & \text{if } h \le h_0, \\ \arg\max_{a\in\mathcal{A}} \tilde{Q}_h^{k,h_0}(s,a), & \text{if } h > h_0. \end{cases} \tag{13}$$

Meanwhile, the lower confidence bounds of the the same value functions are

$$\begin{cases} \underset{\sim}{Q}_h^{k,h_0}(s,a) \triangleq \max\Bigg(0, r_h(s,a) - 3\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} - 22H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\ \qquad - \frac{2}{H}\hat{P}_h^k(\tilde{V}_{h+1}^{k,h_0} - \underset{\sim}{V}_{h+1}^{k,h_0})(s,a) + \hat{P}_h^k\underset{\sim}{V}_{h+1}^{k,h_0}(s,a)\Bigg), \\ \underset{\sim}{V}_h^{k,h_0}(s) \triangleq \langle \pi_h^{k,h_0}(\cdot|s), \tilde{Q}_h^{k,h_0}(s,\cdot)\rangle. \end{cases} \tag{14}$$

The following lemma shows that the above construction are valid UCBs and LCBs.

**Lemma C.5** (UCB and LCB)**.** *With $\tilde{Q}^{k,h_0}$, $\underset{\sim}{Q}^{k,h_0}$, $\tilde{V}^{k,h_0}$, $\underset{\sim}{V}^{k,h_0}$ defined in Equations (12) and (14), the true value functions $Q^{k,h_0}, V^{k,h_0}, Q^{\star,h_0}, V^{\star,h_0}$ can be bounded as:*

$$\underset{\sim}{Q}_h^{k,h_0}(s,a) \overset{(i)}{\le} Q_h^{k,h_0}(s,a) \overset{(ii)}{\le} Q_h^{\star,h_0}(s,a) \overset{(iii)}{\le} \tilde{Q}_h^{k,h_0}(s,a), \tag{15}$$

$$\underset{\sim}{V}_h^{k,h_0}(s) \overset{(iv)}{\le} V_h^{k,h_0}(s) \overset{(v)}{\le} V_h^{\star,h_0}(s) \overset{(vi)}{\le} \tilde{V}_h^{k,h_0}(s). \tag{16}$$

*Proof.* First, we note that due to Lemma C.2, inequalities $(ii)$ and $(v)$ hold for any $h \in [1:H]$.

We then use induction to prove the other four inequalities hold. More specifically, we prove that: **1)** if $(iv)$ and $(vi)$ hold for $h+1, \forall h \in [1:H]$, then $(i)$ and $(iii)$ must hold for $h$, and **2)** if $(i)$ and $(iii)$ hold for any $h \in [1:H]$, then $(iv)$ and $(vi)$ must hold for $h$ as well. For the base case $h = H+1$, all value functions are zeros. Thus, $(iv)$ and $(vi)$ hold for $h = H+1$. We now assume $(iv)$ and $(vi)$ is true for any $h+1$, and prove 1) and 2) recursively through induction.

**Step 1), part 1, inequality** $(iii)$**:** We prove that inequality $(iii)$ in Equation (15) holds for any $h \in [1:H]$. It suffices to consider the case when $\tilde{Q}_h^{k,h_0} < H$. We have

$$\begin{aligned} \tilde{Q}_h^{k,h_0} - Q_h^{\star,h_0} =& 3\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 14H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\ &+ \frac{1}{H}\hat{P}_h^k(\tilde{V}_{h+1}^{k,h_0} - \underset{\sim}{V}_{h+1}^{k,h_0})(s,a) + \hat{P}_h^k\tilde{V}_{h+1}^{k,h_0}(s,a) - P_hV_{h+1}^{\star,h_0}(s,a) \\ =& 3\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 14H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\ &+ \frac{1}{H}\hat{P}_h^k(\tilde{V}_{h+1}^{k,h_0} - \underset{\sim}{V}_{h+1}^{k,h_0})(s,a) + \hat{P}_h^k(\tilde{V}_{h+1}^{k,h_0} - V_{h+1}^{\star,h_0})(s,a) + (\hat{P}_h^k - P_h)V_{h+1}^{\star,h_0}(s,a). \end{aligned} \tag{17}$$

Under good event $\mathcal{E}^\star(V^{\star,h_0},\delta')$, we can bound the last term $(\hat{P}_h^k - P_h)V_{h+1}^{\star,h_0}(s,a)$ as follows:

$$|(\hat{P}_h^k - P_h)V_{h+1}^{\star,h_0}(s,a)| \le \sqrt{2\mathrm{Var}_{P_h}(V_{h+1}^{\star,h_0})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 3H\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}. \tag{18}$$

We further bound the true variance $\mathrm{Var}_{P_h}(V^{\star,h_0}_{h+1})(s,a)$ with the empirical variance $\mathrm{Var}_{\hat{P}_h}(\tilde{V}^{k,h_0}_{h+1})(s,a)$ as follows:

$$\begin{aligned}\mathrm{Var}_{P_h}(V^{\star,h_0}_{h+1})(s,a) &\overset{(a)}{\leq} 2\mathrm{Var}_{\hat{P}_h}(V^{\star,h_0}_{h+1})(s,a) + 4H^2\frac{\beta(n^k_h(s,a),\delta')}{n^k_h(s,a)}\\ &\overset{(b)}{\leq} 4\mathrm{Var}_{\hat{P}_h}(\tilde{V}^{\star,h_0}_{h+1})(s,a) + 4H\hat{P}^k_h|\tilde{V}^{k,h_0}_{h+1} - V^{\star,h_0}_{h+1}|(s,a) + 4H^2\frac{\beta(n^k_h(s,a),\delta')}{n^k_h(s,a)}\\ &\overset{(c)}{\leq} 4\mathrm{Var}_{\hat{P}_h}(\tilde{V}^{\star,h_0}_{h+1})(s,a) + 4H\hat{P}^k_h(\tilde{V}^{k,h_0}_{h+1} - \underset{\sim}{V}^{k,h_0}_{h+1})(s,a) + 4H^2\frac{\beta(n^k_h(s,a),\delta')}{n^k_h(s,a)},\end{aligned}$$

where $(a)$ follows from Lemma F.1 and the definition of good event $\mathcal{E}(\delta/3)$, $(b)$ is due to Lemma F.2, and $(c)$ is due to the induction hypothesis.

Plugging the bound of variance to Equation (18) and applying the facts that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ , $\sqrt{xy} \leq x + y$ and $\beta^\star < \beta$, we have

$$|(\hat{P}^k_h - P_h)V^{\star,h_0}_{h+1}(s,a)| \leq 3\sqrt{\mathrm{Var}_{\hat{P}_h}(\tilde{V}^{k,h_0}_{h+1})(s,a)\frac{\beta^\star(n^k_h(s,a),\delta')}{n^k_h(s,a)}} + \frac{1}{H}\hat{P}^k_h(\tilde{V}^{k,h_0}_{h+1} - \underset{\sim}{V}^{k,h_0}_{h+1})(s,a) + 14H^2\frac{\beta(n^k_h(s,a),\delta')}{n^k_h(s,a)}. \tag{19}$$

Now, plugging Equation (19) back to Equation (17), we have

$$\tilde{Q}^{k,h_0}_h - Q^{\star,h_0}_h \geq \hat{P}^k_h(\tilde{V}^{k,h_0}_{h+1} - V^{\star,h_0}_{h+1})(s,a) \geq 0,$$

where $\tilde{V}^{k,h_0}_{h+1} - V^{\star,h_0}_{h+1} \geq 0$ comes from the induction hypothesis.

**Step 1), part 2, inequality** $(i)$**:** For inequality $(i)$ in Equation (15), it suffices to consider the case when $\underset{\sim}{Q}^{k,h_0}_h > 0$. We have

$$\begin{aligned}Q^{k,h_0}_h(s,a) - \underset{\sim}{Q}^{k,h_0}_h(s,a) =& (P_h - \hat{P}^k_h)V^{\star,h_0}_{h+1}(s,a) + (P_h - \hat{P}^k_h)(\underset{\sim}{V}^{k,h_0}_{h+1} - V^{\star,h_0}_{h+1})(s,a) + P_h(V^{k,h_0}_{h+1} - \underset{\sim}{V}^{k,h_0}_{h+1})(s,a)\\ &+ 3\sqrt{\mathrm{Var}_{\hat{P}^k_h}(\tilde{V}^{k,h_0}_{h+1})(s,a)\frac{\beta^\star(n^k_h(s,a),\delta')}{n^k_h(s,a)}} + 22H^2\frac{\beta(n^k_h(s,a),\delta')}{n^k_h(s,a)}\\ &+ \frac{2}{H}\hat{P}^k_h(\tilde{V}^{k,h_0}_{h+1} - \underset{\sim}{V}^{k,h_0}_{h+1})(s,a).\end{aligned} \tag{20}$$

We have established a bound for $|(P_h - \hat{P}^k_h)V^{\star,h_0}_{h+1}(s,a)|$ in Equation (19). It then suffice to bound $|(P_h - \hat{P}^k_h)(\underset{\sim}{V}^{k,h_0}_{h+1} - V^{\star,h_0}_{h+1})(s,a)|$ as follows.

Because of Lemma F.3, together with good event $\mathcal{E}(\delta/3)$, we have

$$|(P_h - \hat{P}^k_h)(\underset{\sim}{V}^{k,h_0}_{h+1} - V^{\star,h_0}_{h+1})(s,a)| \leq \sqrt{2\mathrm{Var}_{P_h}(V^{\star,h_0}_{h+1} - \underset{\sim}{V}^{k,h_0}_{h+1})(s,a)} + \frac{2}{3}H\frac{\beta(n^k_h(s,a),\delta')}{n^k_h(s,a)}. \tag{21}$$

Moreover, by Lemma F.1,

$$\mathrm{Var}_{P_h}(V^{\star,h_0}_{h+1} - \underset{\sim}{V}^{k,h_0}_{h+1})(s,a) \leq 2\mathrm{Var}_{\hat{P}^k_h}(V^{\star,h_0}_{h+1} - \underset{\sim}{V}^{k,h_0}_{h+1})(s,a) + 4H^2\frac{\beta(n^k_h(s,a),\delta')}{n^k_h(s,a)}. \tag{22}$$

Plugging Equation (22) into Equation (21) and applying $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ and $\sqrt{xy} \leq x + y$, we can bound $|(P_h - \hat{P}^k_h)(\underset{\sim}{V}^{k,h_0}_{h+1} - V^{\star,h_0}_{h+1})(s,a)|$ as follows:

$$|(P_h - \hat{P}^k_h)(\underset{\sim}{V}^{k,h_0}_{h+1} - V^{\star,h_0}_{h+1})(s,a)| \leq \frac{1}{H}\hat{P}^k_h(\tilde{V}^{k,h_0}_{h+1} - \underset{\sim}{V}^{k,h_0}_{h+1})(s,a) + 8H^2\frac{\beta(n^k_h(s,a),\delta')}{n^k_h(s,a)}. \tag{23}$$

Now, plugging Equation (23) and Equation (19) back to Equation (20), we have

$$Q^{k,h_0}_h(s,a) - \underset{\sim}{Q}^{k,h_0}_h(s,a) \geq P_h(V^{k,h_0}_{h+1} - \underset{\sim}{V}^{k,h_0}_{h+1})(s,a) \geq 0,$$

where $V_{h+1}^{k,h_0} - \underset{\sim}{V}_{h+1}^{k,h_0} \geq 0$ comes from the induction hypothesis.

**Step 2):** With all inequalities in Equation (15) being proved, we use them to prove inequalities $(iv)$ and $(vi)$ in Equation (16). Since $\underset{\sim}{V}_h^{k,h_0}(s)$ and $V_h^{k,h_0}(s)$ share the same policy $\pi_h^{k,h_0}$, inequality $(iv)$ can be derived from $\underset{\sim}{Q}_h^{k,h_0}(s,a) \leq Q_h^{k,h_0}(s,a)$. That is,

$$\underset{\sim}{V}_h^{k,h_0}(s) = \langle \pi_h^{k,h_0}(\cdot|s), \underset{\sim}{Q}_h^{k,h_0}(s,\cdot)\rangle \leq \langle \pi_h^{k,h_0}(\cdot|s), Q_h^{k,h_0}(s,\cdot)\rangle = V_h^{k,h_0}(s).$$

To show inequality $(vi)$, we consider two cases: $h > h_0$ and $h \leq h_0$. When $h > h_0$, policy $\pi_h^{k,h_0}$ is the optimistic policy corresponding to $\tilde{Q}_h^{k,h_0}$. Therefore, we have

$$\tilde{V}_h^{k,h_0}(s) = \tilde{Q}_h^{k,h_0}(s, \pi_h^{k,h_0}(s)) \geq \tilde{Q}_h^{k,h_0}(s, \pi_h^{\star}(s)) \geq Q_h^{k,h_0}(s, \pi_h^{\star}(s)) = V_h^{k,h_0}(s).$$

When $h \leq h_0$, both policies $\pi_h^{\star,h_0}$ and $\pi_h^{k,h_0}$ are the baseline policy $\pi_h^b$. Thus, we can use $Q_h^{\star,h_0} \leq \tilde{Q}_h^{k,h_0}$ from Equation (15) to derive $V_h^{\star,h_0}(s) \leq \tilde{V}_h^{k,h_0}(s)$. That is,

$$V_h^{\star,h_0}(s) = \langle \pi_h^b(\cdot|s), Q_h^{\star,h_0}(s,\cdot)\rangle \leq \langle \pi_h^b(\cdot|s), \tilde{Q}_h^{k,h_0}(s,\cdot)\rangle = \tilde{V}_h^{k,h_0}(s).$$

Combining these two cases, we have established $V_h^{\star,h_0}(s) \leq \tilde{V}_h^{k,h_0}(s)$ for any $h$. □

After verifying the validity of UCBs and LCBs in Lemma C.5, we provide the following lemma to quantify the estimation error induced by the lower bound.

**Lemma C.6.** *Define $G_h^{k,h_0}$ as*

$$G_h^{k,h_0}(s,a) = \min\left(H, 6\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a)\frac{\beta^{\star}(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 36H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} + \left(1+\frac{3}{H}\right)\hat{P}_h^k\pi_{h+1}^{k,h_0}G_{h+1}^{k,h_0}(s,a)\right). \tag{24}$$

*Then, the estimation error between $Q_h^{\star,h_0}$, $V_h^{\star,h_0}(s)$ and $\underset{\sim}{Q}_h^{k,h_0}$, $\underset{\sim}{V}_h^{k,h_0}$ can be bounded as*

$$\begin{aligned} Q_h^{\star,h_0}(s,a) - \underset{\sim}{Q}_h^{k,h_0}(s,a) &\leq G_h^{k,h_0}(s,a), \\ V_h^{\star,h_0}(s) - \underset{\sim}{V}_h^{k,h_0}(s) &\leq \langle \pi_h^{k,h_0}(\cdot|s), G_h^{k,h_0}(s,\cdot)\rangle. \end{aligned}$$

*Proof.* $Q_h^{\star,h_0}(s,a) - \underset{\sim}{Q}_h^{k,h_0}(s,a)$ can be directly calculated as follows.

$$\begin{aligned} Q_h^{\star,h_0}(s,a) - \underset{\sim}{Q}_h^{k,h_0}(s,a) &\overset{(a)}{\leq} \tilde{Q}_h^{k,h_0}(s,a) - \underset{\sim}{Q}_h^{k,h_0}(s,a) \\ &\overset{(b)}{\leq} \hat{P}_h^k(\tilde{V}_{h+1}^{k,h_0} - \underset{\sim}{V}_{h+1}^{k,h_0})(s,a) + 6\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a)\frac{\beta^{\star}(n_h^k(s,a),\delta')}{n_h^k(s,a)}} \\ &\quad + 36H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} + \frac{3}{H}\hat{P}_h^k(\tilde{V}_{h+1}^{k,h_0} - \underset{\sim}{V}_{h+1}^{k,h_0})(s,a) \\ &\leq 6\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a)\frac{\beta^{\star}(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 36H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\ &\quad + (1+\frac{3}{H})\hat{P}_h^k(\tilde{V}_{h+1}^{k,h_0} - \underset{\sim}{V}_{h+1}^{k,h_0})(s,a), \end{aligned}$$

where (a) follows from Lemma C.5 and (b) follows from the definitions of $\tilde{Q}_h^{k,h_0}(s,a)$ and $\underset{\sim}{Q}_h^{k,h_0}(s,a)$ in Equation (12) and Equation (14), respectively.

Then, following the same argument, for $V$ functions, we have

$$V_h^{\star,h_0}(s) - \underset{\sim}{V}_h^{k,h_0}(s) \leq \tilde{V}_h^{k,h_0}(s) - \underset{\sim}{V}_h^{k,h_0}(s) \leq \langle \pi_h^{k,h_0}(\cdot|s), (\tilde{Q}_h^{k,h_0} - \underset{\sim}{Q}_h^{k,h_0})(s,\cdot)\rangle.$$

Combining the above two inequalities with the definition of $G_h^{k,h_0}$, we have

$$Q_h^{\star,h_0}(s,a) - \underset{\sim}{Q}_h^{k,h_0}(s,a) \le G_h^{k,h_0}(s,a),$$
$$V_h^{\star,h_0}(s) - \underset{\sim}{V}_h^{k,h_0}(s) \le \langle \pi_h^{k,h_0}(\cdot|s), G_h^{k,h_0}(s,\cdot)\rangle,$$

which completes the proof. □

In the following lemma, we aim to upper bound $\pi_1^{k,h_0} G_1^{k,h_0}$.

**Lemma C.7** (Upper bound $\pi_1^{k,h_0} G_1^{k,h_0}$)**.** *For any $k$ and $h_0$, we have*

$$\pi_1^{k,h_0} G_1^{k,h_0}(s_1) \le 24e^{13} \sum_{h=1}^{H} \sum_{s,a} d_h^{k,h_0}(s,a) \sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0})(s,a) \left( \frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1} \right)}$$
$$+ 336e^{13} H^2 \sum_{h=1}^{H} \sum_{s,a} d_h^{k,h_0}(s,a) \left( \frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1} \right),$$

*where $G_h^{k,h_0}$ is defined in Equation (24) and $d_h^{k,h_0}$ is the occupancy measure under policy $\pi^{k,h_0}$.*

*Proof.* From the definition of $G_h^{k,h_0}$ in Equation (24), we have

$$G_h^{k,h_0}(s,a) \le 6 \sqrt{\underbrace{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a)}_{\text{(I)}} \frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 36H^2 \frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} + \left(1 + \frac{3}{H}\right) \underbrace{\hat{P}_h^k \pi_{h+1}^{k,h_0} G_{h+1}^{k,h_0}(s,a)}_{\text{(II)}}. \tag{25}$$

In order to bound term (II), we use Lemma F.3 and the fact that $\sqrt{xy} \le x + y$ to obtain

$$\begin{aligned}(\hat{P}_h^k - P_h)\pi_{h+1}^k G_{h+1}^{k,h_0}(s,a) &\le \sqrt{2\mathrm{Var}_{P_h}(\pi_{h+1}^k G_{h+1}^{k,h_0})(s,a) \frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + \frac{2}{3} H \frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\ &\le \frac{1}{H} P_h \pi_{h+1}^k G_{h+1}^{k,h_0}(s,a) + 3H^2 \frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)}.\end{aligned} \tag{26}$$

For term (I), we have

$$\begin{aligned}\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a) &\overset{(a)}{\le} 2\mathrm{Var}_{P_h}(\tilde{V}_{h+1}^{k,h_0})(s,a) + 4H^2 \frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\ &\overset{(b)}{\le} 4\mathrm{Var}_{P_h}(V_{h+1}^{\star,h_0})(s,a) + 4HP_h|\tilde{V}_{h+1}^{k,h_0} - V_{h+1}^{\star,h_0}|(s,a) + 4H^2 \frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\ &\overset{(c)}{\le} 4\mathrm{Var}_{\hat{P}_h}(\tilde{V}_{h+1}^{\star,h_0})(s,a) + 4H\hat{P}_h^k(\tilde{V}_{h+1}^{k,h_0} - \underset{\sim}{V}_{h+1}^{k,h_0})(s,a) + 4H^2 \frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)},\end{aligned}$$

where $(a)$ follows from Lemma F.1, $(b)$ follows from Lemma F.2, and $(c)$ follows from Lemma C.5.

Moreover, we can use $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$, $\sqrt{xy} \le x + y$ to obtain

$$\begin{aligned}&\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a) \frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} \\ &\le 2\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0})(s,a) \frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 6H^2 \frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} + \frac{1}{H} P_h \pi_{h+1}^{k,h_0} G_{h+1}^{k,h_0}(s,a).\end{aligned} \tag{27}$$

Applying Equation (26) and Equation (27) to Equation (25), we have

$$
\begin{aligned}
&G_h^{k,h_0}(s,a) \\
&\quad\leq 6\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,h_0})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 36H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} + \left(1+\frac{3}{H}\right)\hat{P}_h^k\pi_{h+1}^{k,h_0}G_{h+1}^{k,h_0}(s,a) \\
&\quad\leq 12\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 36H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} + \frac{6}{H}P_h\pi_{h+1}^{k,h_0}G_{h+1}^{k,h_0}(s,a) \\
&\qquad + 36H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} + \left(1+\frac{3}{H}\right)P_h\pi_{h+1}^{k,h_0}G_{h+1}^{k,h_0}(s,a) \\
&\qquad + \left(1+\frac{3}{H}\right)\left(\frac{1}{H}P_h\pi_{h+1}^{k}G_{h+1}^{k,h_0}(s,a) + 3H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)}\right) \\
&\quad\leq 12\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 84H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} + \left(1+\frac{13}{H}\right)P_h\pi_{h+1}^{k}G_{h+1}^{k,h_0}(s,a).
\end{aligned}
$$

In addition, since $G_h^{k,h_0}$ is upper bounded by $H$ by definition, and $\beta(n_h^k(s,a),\delta') > \beta^\star(n_h^k(s,a),\delta')$, we have

$$
\begin{aligned}
&G_h^{k,h_0}(s,a) \\
&\quad\leq \min\left\{12\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 84H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} + \left(1+\frac{13}{H}\right)P_h\pi_{h+1}^{k}G_{h+1}^{k,h_0}(s,a), H\right\} \\
&\quad\leq 12\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0})(s,a)\left(\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}\wedge 1\right)} + 84H^2\left(\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)}\wedge 1\right) + \left(1+\frac{13}{H}\right)P_h\pi_{h+1}^{k}G_{h+1}^{k,h_0}(s,a).
\end{aligned}
$$

Using $(1+\frac{13}{H})^H \leq e^{13}$ and unfolding the above inequality, we have

$$
\begin{aligned}
\pi_1^{k,h_0}G_1^{k,h_0}(s_1) \leq& 12e^{13}\sum_{h=1}^{H}\sum_{s,a}d_h^{k,h_0}(s,a)\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0})(s,a)\left(\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}\wedge 1\right)} \\
&+ 84e^{13}H^2\sum_{h=1}^{H}\sum_{s,a}d_h^{k,h_0}(s,a)\left(\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)}\wedge 1\right).
\end{aligned}
$$

Finally, we use Lemma F.5 to transform $n_h^k(s,a)$ to $\bar{n}_h^k(s,a)$ and obtain

$$
\begin{aligned}
\pi_1^{k,h_0}G_1^{k,h_0}(s_1) \leq& 24e^{13}\sum_{h=1}^{H}\sum_{s,a}d_h^{k,h_0}(s,a)\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0})(s,a)\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)} \\
&+ 336e^{13}H^2\sum_{h=1}^{H}\sum_{s,a}d_h^{k,h_0}(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right),
\end{aligned}
$$

which completes the proof. □

### C.3. Step Three: Finite Episodes for Step Mixture Policies

In the previous section, we have characterized the UCBs and LCBs for the step-wise optimistic policies ($\pi^{k,h_0}$) and bound the corresponding estimation errors. In this step, we extend the result for $\pi^{k,h_0}$ to step mixture policies $\pi^k$ and prove that the number of the episodes of which the executed policy $\pi^k$ is not equal to the optimistic policy $\bar{\pi}^k$ (i.e., $\pi^{k,0}$), is finite under the StepMix algorithm. We refer to this result as the *finite non-optimistic policy lemma*.

To establish the finite non-optimistic policy lemma, we first extend the result of Lemma C.7 from $\pi^{k,h_0}$ to step mixture policies.

**Lemma C.8.** *For a step mixture policy $\pi^k$ mixed from two policies $\pi^{k,h_0}$ and $\pi^{k,h_0-1}$, denoted as $\pi^k = (1-\rho)\pi^{k,h_0} + \rho\pi^{k,h_0-1}$ for some $\rho \in (0,1)$, the following inequality holds:*

$$\begin{aligned}&(1-\rho)\pi_1^{k,h_0}G_1^{k,h_0}(s_1) + \rho\pi_1^{k,h_0-1}G_1^{k,h_0-1}(s_1)\\ &\quad\le 24e^{13}H\sqrt{\sum_{h=1}^{H}\sum_{s,a} d_h^k(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)} + 336e^{13}H^2\sum_{h=1}^{H}\sum_{s,a} d_h^k(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right),\end{aligned} \tag{28}$$

*where $d_h^k(s,a)$ is the occupancy measure under policy $\pi^k$ and $G_h^{k,h_0}$ is defined in Equation (24).*

*Proof.* Since $\pi^k = (1-\rho)\pi^{k,h_0} + \rho\pi^{k,h_0-1}$ is the step mixture policy mixed from two policies that differ at only one step, by Lemma C.3, the occupancy measure under $\pi^k$ satisfies

$$d_h^k(s,a) = (1-\rho)d^{k,h_0}(s,a) + \rho d^{k,h_0-1}(s,a).$$

Using Lemma C.7, we have

$$\begin{aligned}&(1-\rho)\pi_1^{k,h_0}G_1^{k,h_0}(s_1) + \rho\pi_1^{k,h_0-1}G_1^{k,h_0-1}(s_1)\\ &\le\rho\Bigg(24e^{13}\sum_{h=1}^{H}\sum_{s,a} d_h^{k,h_0-1}(s,a)\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0-1})(s,a)\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)}\\ &\qquad + 336e^{13}H^2\sum_{h=1}^{H}\sum_{s,a} d_h^{k,h_0-1}(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)\Bigg)\\ &\quad + (1-\rho)\Bigg(24e^{13}\sum_{h=1}^{H}\sum_{s,a} d_h^{k,h_0}(s,a)\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0})(s,a)\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)}\\ &\qquad + 336e^{13}H^2\sum_{h=1}^{H}\sum_{s,a} d_h^{k,h_0}(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)\Bigg).\end{aligned} \tag{29}$$

It is worth noting that

$$\begin{aligned}&\sum_{h=1}^{H}\sum_{s,a} d_h^k(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)\\ &\quad= \rho\sum_{h=1}^{H}\sum_{s,a} d_h^{k,h_0-1}(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right) + (1-\rho)\sum_{h=1}^{H}\sum_{s,a} d_h^{k,h_0}(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right).\end{aligned} \tag{30}$$

Thus, to prove Equation (28), it suffices to show that

$$\begin{aligned}&\sum_{h=1}^{H}\sum_{s,a} \rho d_h^{k,h_0-1}(s,a)\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0-1})(s,a)\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)}\\ &\qquad + \sum_{h=1}^{H}\sum_{s,a}(1-\rho)d_h^{k,h_0}(s,a)\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0})(s,a)\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)}\\ &\le H\sqrt{\sum_{h=1}^{H}\sum_{s,a} d_h^k(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)}.\end{aligned} \tag{31}$$

Due to the Cauchy's inequality, we have

$$\text{LHS of (31)} \leq \sqrt{\sum_{h=1}^{H}\sum_{s,a}\rho d_h^{k,h_0-1}(s,a)\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0-1})(s,a)+(1-\rho)d_h^{k,h_0}(s,a)\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0})(s,a)} \times\sqrt{\sum_{h=1}^{H}\sum_{s,a}(\rho d_h^{k,h_0-1}(s,a)+(1-\rho)d_h^{k,h_0}(s,a))\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)}. \tag{32}$$

Besides, due to Lemma F.4, we have

$$\sum_{h=1}^{H}\sum_{s,a}d_h^{k,h_0}(s,a)\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0})(s,a)\leq \mathbb{E}_{\pi^{k,h_0}}\left[\left(\sum_{h=1}^{H}r_h(s_h,a_h)-V_1^{k,h_0}(s_1)\right)^2\right]\leq H^2.$$

Similarly, we also have $\sum_{h=1}^{H}\sum_{s,a}d_h^{k,h_0-1}(s,a)\mathrm{Var}_{P_h}(V_{h+1}^{k,h_0-1})(s,a)\leq H^2$. Together with Equation (30), we have

$$\text{RHS of (32)}\leq H\sqrt{\sum_{h=1}^{H}\sum_{s,a}d_h^k(s,a)\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)},$$

which completes the proof. □

Equipped with Lemma C.8, we are ready to establish the *finite non-optimistic policy lemma*, which states that for step mixture policies, there are only finite episodes in which $\pi^k\neq\pi^{k,0}$.

**Lemma C.9** (Finite non-optimistic policy lemma)**.** *Define $\mathcal{N}=\{k|k\in[K],\pi^k\neq\pi^{k,0}\}$. Then, the cardinality of $\mathcal{N}$ is upper bounded by*

$$|\mathcal{N}|\leq\left(\frac{4608e^{26}}{\kappa^2}+\frac{2688e^{13}S}{\kappa}\right)H^3SA\log^2(K+1)=\tilde{O}\left(\left(\frac{1}{\kappa^2}+\frac{S}{\kappa}\right)H^3SA\right),$$

*where $\kappa=V_1^{\pi^b}-\gamma$.*

*Proof.* By the definition of $\mathcal{N}$, we have

$$\begin{aligned}
|\mathcal{N}|\kappa &\overset{(a)}{\leq}\sum_{k\in\mathcal{N}}\left(V_1^{\pi^b}-(\rho\underline{V}_1^{k,h_0-1}+(1-\rho)\underline{V}_1^{k,h_0})\right) && (33)\\
&\overset{(b)}{\leq}\sum_{k\in\mathcal{N}}\left((\rho V_1^{\star,h_0-1}+(1-\rho)V_1^{\star,h_0})-(\rho\underline{V}_1^{k,h_0-1}+(1-\rho)\underline{V}_1^{k,h_0})\right)\\
&=\sum_{k\in\mathcal{N}}\left(\rho(V_1^{\star,h_0-1}-\underline{V}_1^{k,h_0-1})+(1-\rho)(V_1^{\star,h_0}-\underline{V}_1^{k,h_0})\right)\\
&\overset{(c)}{\leq}\sum_{k\in\mathcal{N}}\left(\rho\pi_1^{k,h_0-1}G_1^{k,h_0-1}+(1-\rho)\pi_1^{k,h_0}G_1^{k,h_0}\right) && (34)\\
&\overset{(d)}{\leq}24e^{13}H\sum_{k\in\mathcal{N}}\sqrt{\sum_{h=1}^{H}\sum_{s,a}d_h^k(s,a)\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)}\\
&\quad+336e^{13}H^2\sum_{k\in\mathcal{N}}\sum_{h=1}^{H}\sum_{s,a}d_h^k(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)\\
&\overset{(e)}{\leq}24e^{13}H\sqrt{|\mathcal{N}|}\sqrt{\sum_{k\in\mathcal{N}}\sum_{h=1}^{H}\sum_{s,a}d_h^k(s,a)\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)}
\end{aligned}$$

$$+ 336e^{13}H^2 \sum_{k\in\mathcal{N}} \sum_{h=1}^{H} \sum_{s,a} d_h^k(s,a) \left( \frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1} \right), \tag{35}$$

where $(b)$ follows from Lemma C.2, $(c)$ is due to Lemma C.6, $(d)$ follows from Lemma C.8, $(e)$ is due to the Cauchy's inequality. For inequality $(a)$, by the design of StepMix, when $\pi^k \neq \pi^{k,0}$, we must have $\rho \underset{\sim}{V}_1^{k,h_0-1} + (1-\rho)\underset{\sim}{V}_1^{k,h_0} = \gamma$ or $\rho \underset{\sim}{V}_1^{k,h_0-1} + (1-\rho)\underset{\sim}{V}_1^{k,h_0} = \underset{\sim}{V}_1^{\pi^{k,H}} \leq \gamma$. Both cases indicate that inequality $(a)$ holds.

We can also bound the summation as follows:

$$\begin{aligned}
\sum_{k\in\mathcal{N}} \sum_{h=1}^{H} \sum_{s,a} d_h^k(s,a) \left( \frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1} \right) \leq & \sum_{h=1}^{H} \sum_{s,a} \sum_{k\in\mathcal{N}} d_h^k(s,a) \left( \frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1} \right) \\
\leq & \beta^\star(K,\delta') \sum_{h=1}^{H} \sum_{s,a} \sum_{k\in\mathcal{N}} \left( \frac{d_h^k(s,a)}{\bar{n}_h^k(s,a) \vee 1} \right) \\
\overset{(a)}{\leq} & \beta^\star(K,\delta') \sum_{h=1}^{H} \sum_{s,a} 4 \log\left(|\mathcal{N}|+1\right) \\
\leq & 4HSA\beta^\star(K,\delta') \log\left(|\mathcal{N}|+1\right),
\end{aligned} \tag{36}$$

where inequality (a) follows from Lemma F.6.

Similar to Equation (36), we also have

$$\sum_{k\in\mathcal{N}} \sum_{h=1}^{H} \sum_{s,a} d_h^k(s,a) \left( \frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1} \right) \leq 4HSA\beta(K,\delta') \log\left(|\mathcal{N}|+1\right). \tag{37}$$

Hence, putting the bound of summations into Equation (35), we have

$$|\mathcal{N}|\kappa \leq 48e^{13}\sqrt{|\mathcal{N}|}\sqrt{H^3SA\beta^\star(K,\delta')\log\left(|\mathcal{N}|+1\right)} + 1344e^{13}H^3SA\beta(K,\delta')\log\left(|\mathcal{N}|+1\right).$$

Rearranging the terms, we conclude that $|\mathcal{N}| \leq (\frac{4608e^{26}}{\kappa^2} + \frac{2688e^{13}S}{\kappa})H^3SA\log^2(K+1) = \tilde{O}\left((\frac{1}{\kappa^2} + \frac{S}{\kappa})H^3SA\right)$, which completes the proof. □

### C.4. Step Four: Putting Everything Together

Finally, with all the results above, we can prove Theorem 4.2.

**Theorem C.10** (The complete version of Theorem 4.2)**.** *Given $\delta \in (0,1)$, set $\delta' = \frac{\delta}{3(H+1)}$, $\beta = \log(SAH/\delta') + S\log(8e(K+1))$, and $\beta^\star = \log(SAH/\delta') + \log(8e(K+1))$. Then, with probability at least $1-\delta$, StepMix satisfies constraint in (2) and achieves a regret upper bounded as*

$$Reg(K) \leq \tilde{O}\left(\sqrt{H^3SAK} + H^3S^2A + H^3SA\Delta_0\left(\frac{1}{\kappa^2} + \frac{S}{\kappa}\right)\right),$$

*where $\Delta_0 = V_1^\star - V_1^{\pi^b}$ and $\kappa = V_1^{\pi^b} - \gamma$.*

*Proof.* We use the same notations specified in Appendix B. Then, under the previously defined good events (which occur with probability at least $1-\delta$), we have

$$\begin{aligned}
\text{Reg}(K) = & \sum_{k=1}^{K} (V_1^\star - V_1^{\pi^k}) \\
\overset{(a)}{=} & \sum_{k\notin\mathcal{N}} (V^{\star,0} - V^{k,0}) + \sum_{k\in\mathcal{N}} (V^\star - V^{\pi^b}) + \sum_{k\in\mathcal{N}} (V^{\pi^b} - V^{\pi^k})
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\leq} \sum_{k\notin\mathcal{N}} (V^{\star,0} - \underset{\sim}{V}^{k,0}) + |\mathcal{N}|\Delta_0 + \sum_{k\in\mathcal{N}} (\rho\pi_1^{k,h_0-1}G_1^{k,h_0-1} + (1-\rho)\pi_1^{k,h_0}G_1^{k,h_0}) \\
&\overset{(c)}{\leq} \sum_{k=1}^{K} (\rho\pi_1^{k,h_0-1}G_1^{k,h_0-1} + (1-\rho)\pi_1^{k,h_0}G_1^{k,h_0}) + |\mathcal{N}|\Delta_0 \\
&\overset{(d)}{\leq} \sum_{k=1}^{K} \Bigg( 24e^{13}H \sqrt{\sum_{h=1}^{H}\sum_{s,a} d_h^k(s,a) \left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)} \\
&\qquad + 336e^{13}H^2 \sum_{h=1}^{H}\sum_{s,a} d_h^k(s,a) \left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)\Bigg) + |\mathcal{N}|\Delta_0,
\end{aligned} \tag{38}
$$

where $(a)$ is due to that $\pi^{\star,0} = \pi^\star$ and $\pi^k = \pi^{k,0}$ when $k \notin \mathcal{N}$; $(b)$ follows from the fact that $\underset{\sim}{V}_1^{k,0}$ is the LCB of $V_1^{k,0}$ and Equation (33) to Equation (34) in the proof of Lemma C.9; $(c)$ is due to $V_1^{\star,0} - \underset{\sim}{V}_1^{k,0} \leq \pi_1^{k,0}G_1^{k,0}$ as shown in Lemma C.6, and $(d)$ is from Lemma C.8.

Using Equations (36) and (37) from the proof of Lemma C.9, we obtain

$$
\begin{aligned}
&\sum_{k=1}^{K} \Bigg( 24e^{13}H \sqrt{\sum_{h=1}^{H}\sum_{s,a} d_h^k(s,a) \left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta)}{\bar{n}_h^k(s,a)\vee 1}\right)} + 336e^{13}H^2 \sum_{h=1}^{H}\sum_{s,a} d_h^k(s,a) \left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)\Bigg) \\
&\quad \leq 48e^{13}\sqrt{H^3SAK\beta^\star(K,\delta')\log(K+1)} + 1344e^{13}H^3SA\beta(K,\delta')\log(K+1).
\end{aligned} \tag{39}
$$

Plugging (39) and the result of Lemma C.9 into Equation (38), we further have:

$$
\begin{aligned}
\mathrm{Reg}(K) &\leq 48e^{13}\sqrt{H^3SAK\beta^\star(K,\delta')\log(K+1)} + 1344e^{13}H^3SA\beta(K,\delta')\log(K+1) \\
&\quad + \left(\frac{4608e^{26}}{\kappa^2} + \frac{2688e^{13}S}{\kappa}\right) H^3SA\log(K+1)\Delta_0 \\
&= \tilde{O}\left(\sqrt{H^3SAK} + H^3S^2A + H^3SA\Delta_0\left(\frac{1}{\kappa^2} + \frac{S}{\kappa}\right)\right).
\end{aligned}
$$

Finally, we prove that StepMix satisfies the constraint episodically. Specifically, for any online policy $\pi^k$, we have

$$
V_1^{\pi^k} = \rho V_1^{k,h_0-1} + (1-\rho)V_1^{k,h_0} \geq \rho\underset{\sim}{V}_1^{k,h_0-1} + (1-\rho)\underset{\sim}{V}_1^{k,h_0}.
$$

If $\pi^k = \pi^{k,H} = \pi^b$, $\pi^k$ must satisfy the constraint, because $\pi^b$ is assumed to be safe. Otherwise, if $h_0 = 0$, that means $\pi^k = \pi^{k,0}$ and $\underset{\sim}{V}_1^{k,0} > \gamma$, so $\pi^k$ must be safe; if $h_0 \neq 0$, we have $\underset{\sim}{V}_1^{k,h_0} \geq \gamma$ and $\underset{\sim}{V}_1^{k,h_0-1} < \gamma$, so that $\rho = \frac{\underset{\sim}{V}_1^{k,h_0}(s_1)-\gamma}{\underset{\sim}{V}_1^{k,h_0}(s_1)-\underset{\sim}{V}_1^{k,h_0-1}}$ guarantees that $\gamma = \rho\underset{\sim}{V}_1^{k,h_0-1} + (1-\rho)\underset{\sim}{V}_1^{k,h_0}$. Since $\rho\underset{\sim}{V}_1^{k,h_0-1} + (1-\rho)\underset{\sim}{V}_1^{k,h_0} \leq \rho V_1^{k,h_0-1} + (1-\rho)V_1^{k,h_0} = V_1^{\pi^k}$, $\pi^k$ is also safe. □

We remark that when $\gamma = 0$ the additive term $\tilde{O}\left(H^3SA\Delta_0\left(\frac{1}{\kappa^2} + \frac{S}{\kappa}\right)\right)$ in Theorem 4.2 can be dropped, as formally stated in the following corollary.

**Corollary C.11** (Vanishing additive term)**.** *When $\gamma = 0$, with all the parameters specified in Theorem 4.2, StepMix satisfies constraint (2) and achieves a regret upper bounded as*

$$
\mathit{Reg}(K) \leq \tilde{O}\left(\sqrt{H^3SAK} + H^3S^2A\right),
$$

*where $\Delta_0 = V_1^\star - V_1^{\pi^b}$, $\kappa = V_1^{\pi^b} - \gamma$.*

*Proof.* If $\gamma = 0$, based on the definition of $\underset{\sim}{Q}_h^{k,h_0}$ in Equation (14), we have $\underset{\sim}{Q}_h^{k,h_0} \geq 0 = \gamma$ for any $k$, $h_0$ and $h$. Thus, under StepMix, the executed policy $\pi^{\tilde{k}}$ must be the optimistic policy $\bar{\pi}^{\tilde{k}}$. Recall that the definition of $\mathcal{N}$ is $\mathcal{N} = \{k | k \in [K], \pi^k \neq \bar{\pi}^k\}$. Therefore, we have $\mathcal{N} = \emptyset$ and $|\mathcal{N}| = 0$.

With $|\mathcal{N}| = 0$ and Equation (38), we have

$$\begin{aligned}
\mathrm{Reg}(K) &\leq \sum_{k=1}^{K}\left(24e^{13}H\sqrt{\sum_{h=1}^{H}\sum_{s,a}d_h^k(s,a)\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)} + 336e^{13}H^2\sum_{h=1}^{H}\sum_{s,a}d_h^k(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)\right) \\
&\quad + |\mathcal{N}|\Delta_0 \\
&= \sum_{k=1}^{K}\left(24e^{13}H\sqrt{\sum_{h=1}^{H}\sum_{s,a}d_h^k(s,a)\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)} + 336e^{13}H^2\sum_{h=1}^{H}\sum_{s,a}d_h^k(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)\right) \\
&\overset{(a)}{\leq} 48e^{13}\sqrt{H^3SAK\beta^\star(K,\delta')\log(K+1)} + 1344e^{13}H^3SA\beta(K,\delta')\log(K+1) \\
&= \tilde{O}\left(\sqrt{H^3SAK} + H^3S^2A\right),
\end{aligned}$$

where (a) is due to Equations (36) and (37). □

## D. Algorithm Design and Analysis of EpsMix Algorithm

In this section, we present the detailed design and analysis of the EpsMix algorithm.

### D.1. Algorithm Design

The EpsMix algorithm is presented in Algorithm 4. The update rule of $\tilde{Q}_h^k$, $\underset{\sim}{Q}_h^k$ is given below, where $\delta' = \delta/4$.

$$\begin{aligned}
\tilde{Q}_h^k(s,a) &\triangleq \min\Bigg(H, r_h(s,a) + 3\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^k)(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 14H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\
&\qquad + \frac{1}{H}\hat{P}_h^k(\tilde{V}_{h+1}^k - \underset{\sim}{V}_{h+1}^k)(s,a) + \hat{P}_h^k\tilde{V}_{h+1}^k(s,a)\Bigg), \\
\underset{\sim}{Q}_h^k(s,a) &\triangleq \max\Bigg(0, r_h(s,a) - 3\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^k)(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} - 22H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\
&\qquad - \frac{2}{H}\hat{P}_h^k(\tilde{V}_{h+1}^k - \underset{\sim}{V}_{h+1}^k)(s,a) + \hat{P}_h^k\underset{\sim}{V}_{h+1}^k(s,a)\Bigg).
\end{aligned} \tag{40}$$

Similarly, the update rule of $\tilde{Q}_h^{k,b}$ and $\underset{\sim}{Q}_h^{k,b}$ is defined as

$$\begin{aligned}
\tilde{Q}_h^{k,b}(s,a) &\triangleq \min\Bigg(H, r_h(s,a) + 3\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,b})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 14H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\
&\qquad + \frac{1}{H}\hat{P}_h^k(\tilde{V}_{h+1}^{k,b} - \underset{\sim}{V}_{h+1}^{k,b})(s,a) + \hat{P}_h^k\tilde{V}_{h+1}^{k,b}(s,a)\Bigg), \\
\underset{\sim}{Q}_h^{k,b}(s,a) &\triangleq \max\Bigg(0, r_h(s,a) - 3\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,b})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} - 22H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \\
&\qquad - \frac{2}{H}\hat{P}_h^k(\tilde{V}_{h+1}^{k,b} - \underset{\sim}{V}_{h+1}^{k,b})(s,a) + \hat{P}_h^k\underset{\sim}{V}_{h+1}^{k,b}(s,a)\Bigg).
\end{aligned} \tag{41}$$

### D.2. Theoretical Analysis

The performance of the EpsMix Algorithm is characterized in the following theorem.

**Theorem D.1** (Regret of EpsMix)**.** *Given $\delta \in (0,1)$, set $\delta' = \frac{\delta}{4}$, $\beta = \log(SAH/\delta') + S\log(8e(K+1))$, and $\beta^\star = \log(SAH/\delta') + \log(8e(K+1))$. Then, with probability at least $1-\delta$, EpsMix (Algorithm 4) simultaneously (i) satisfies*

**Algorithm 4** The EpsMix Algorithm

**Input:** $\pi^b$, $\gamma$, $\beta$, $\beta^\star$, $\mathcal{D}_0 = \emptyset$
**for** $k$ = 1 to $K$ **do**
  Update the model estimate

$$\hat{P}_h^k(s'|s,a) = \begin{cases} n_h^k(s,a,s')/n_h^k(s,a), & \text{if } n_h^k(s,a) > 0, \\ 1/S, & \text{if } n_h^k(s,a) = 0. \end{cases}$$

  *# Optimistic policy identification*
  $\tilde{Q}_{H+1}^k = \underset{\sim}{Q}_{H+1}^k = 0.$
  **for** $h$ = $H$ to 1 **do**
    Update $\tilde{Q}_h^k(s,a), \underset{\sim}{Q}_h^k(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$ according to Equation (40).
    $\bar{\pi}_h^k(s) \leftarrow \arg\max_a \tilde{Q}_h^k(s,a)$, $\tilde{V}_h^k(s) \leftarrow \tilde{Q}_h^k(s, \bar{\pi}_h^k(s))$, $\underset{\sim}{V}_h^k(s) \leftarrow \underset{\sim}{Q}_h^k(s, \bar{\pi}_h^k(s)), \forall s \in \mathcal{S}$.
  **end for**
  *# Evaluate the baseline policy*
  $\tilde{Q}_{H+1}^{k,b} = \underset{\sim}{Q}_{H+1}^{k,b} = 0.$
  **for** $h$ = $H$ to 1 **do**
    Update $\tilde{Q}_h^{k,b}(s,a), \underset{\sim}{Q}_h^{k,b}(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$ according to Equation (41).
    $\tilde{V}_h^{k,b}(s) \leftarrow \tilde{Q}_h^{k,b}(s, \pi_h^b(s))$, $\underset{\sim}{V}_h^{k,b}(s) \leftarrow \underset{\sim}{Q}_h^{k,b}(s, \pi_h^b(s)), \forall s \in \mathcal{S}$.
  **end for**
  *# Safe exploration policy selection*
  **if** $\underset{\sim}{V}_1^k \geq \gamma$ **then**
    $\pi^k = \bar{\pi}^k$.
  **else if** $\underset{\sim}{V}_1^{k,b} < \gamma$ **then**
    $\pi^k = \pi^b$.
  **else**
    $\rho = \frac{\underset{\sim}{V}_1^{k,b}(s_1) - \gamma}{\underset{\sim}{V}_1^{k,b}(s_1) - \underset{\sim}{V}_1^k}$,
    $\pi^k = \rho \bar{\pi}^k \oplus (1-\rho)\pi^b$.
  **end if**
  Execute $\pi^k$ and collect $\{(s_h^k, a_h^k, s_{h+1}^k)\}_{h=1}^H$.
  $\mathcal{D}_n \leftarrow \mathcal{D}_{n-1} \cup \{(s_h^k, a_h^k, s_{h+1}^k)\}_{h=1}^H$.
**end for**

*the conservative constraint in (2), and (ii) achieves a total regret that is upper bounded by*

$$\tilde{O}\left(\sqrt{H^3SAK} + H^3S^2A + H^3SA\Delta_0\left(\frac{1}{\kappa^2} + \frac{S}{\kappa}\right)\right),$$

*where* $\Delta_0 = V_1^\star - V_1^{\pi^b}$ *is the suboptimality gap of the baseline policy, and* $\kappa = V_1^{\pi^b} - \gamma$ *is the tolerable value loss from the baseline policy.*

Before we proceed to prove Theorem D.1, we sketch the proof as follows: First, we establish the UCBs and LCBs of the value functions for the baseline policy $\pi^b$ and the optimal policy $\pi^\star$ in each episode, following similar approaches as in the proof of Theorem 4.2. We then show that the total number of episodes where the algorithm executes $\pi^b$ or the episodic mixture policy is bounded, which ensures that the performance degradation compared with BPI-UCBVI (Ménard et al., 2021) is bounded. Finally, the established LCBs ensure that the conservative constraint is satisfied in each episode.

**Lemma D.2.** *With probability at least* $1-\delta$*, the following good events occur simultaneously:*

$$\mathcal{E}(\delta'), \mathcal{E}^{cnt}(\delta'), \mathcal{E}^\star(V^\star, \delta'), \mathcal{E}^\star(V^{\pi^b}, \delta'),$$

*where* $\delta' = \delta/4$.

*Proof.* This result can be obtained by noting that each of those events hold with probability at least $1-\delta/4$ under Theorem F.7, Theorem F.8 and Theorem F.9, and then taking the union bound. □

In the following proof of EpsMix, we set $\delta' = \delta/4$. We note that Equation (40) and Equation (41) are defined in a similar form as Equation (8). As a result, Lemmas C.5 to C.7 can be directly extended for EpsMix, as stated below. We note that $Q^{k,h_0}$ need to be bounded for every $h_0 \in [H] \cup \{0\}$ in StepMix, while in EpsMix, we only need to bound $Q^k$ and $Q^{k,b}$.

**Lemma D.3** (UCB and LCB for EpsMix)**.** *The relationship between $\tilde{Q}_h^k$, $\underset{\sim}{Q}_h^k$, $\tilde{V}_h^k$, $\underset{\sim}{V}_h^k$ and the corresponding true value functions $Q_h^k$, $V_h^k$, $Q_h^\star$, $V_h^\star$ are specified in the following inequalities:*

$$\underset{\sim}{Q}_h^k(s,a) \le Q_h^k(s,a) \le Q_h^\star(s,a) \le \tilde{Q}_h^k(s,a),$$
$$\underset{\sim}{V}_h^k(s) \le V_h^k(s) \le V_h^\star(s) \le \tilde{V}_h^k(s),$$

*In addition, the relationships between $\tilde{Q}_h^{k,b}$, $\underset{\sim}{Q}_h^{k,b}$, $\tilde{V}_h^{k,b}$, $\underset{\sim}{V}_h^{k,b}$, and the true value functions $Q_h^{\pi^b}, V_h^{\pi^b}$ are specified in the following inequalities:*

$$\underset{\sim}{Q}_h^{k,b}(s,a) \le Q_h^{\pi^b}(s,a) \le \tilde{Q}_h^{k,b}(s,a),$$
$$\underset{\sim}{V}_h^{k,b}(s) \le V_h^{\pi^b}(s) \le \tilde{V}_h^{k,b}(s).$$

**Lemma D.4.** *Define $G_h^k$ and $G_h^{k,b}$ as*

$$G_h^k(s,a) = \min\left(H, 6\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^k)(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 36H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} + (1+\frac{3}{H})\hat{P}_h^k\pi_{h+1}^k G_{h+1}^k(s,a)\right), \tag{42}$$

$$G_h^{k,b}(s,a) = \min\left(H, 6\sqrt{\mathrm{Var}_{\hat{P}_h^k}(\tilde{V}_{h+1}^{k,b})(s,a)\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)}} + 36H^2\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} + (1+\frac{3}{H})\hat{P}_h^k\pi_{h+1}^{k,b} G_{h+1}^{k,b}(s,a)\right). \tag{43}$$

*Then, the estimation error between $Q_h^\star$, $V_h^\star$ and $\underset{\sim}{Q}_h^k$, $\underset{\sim}{V}_h^k$ can be bounded as*

$$Q_h^\star(s,a) - \underset{\sim}{Q}_h^k(s,a) \le G_h^k(s,a),$$

$$V_h^\star(s) - \underset{\sim}{V}_h^k(s) \le \langle \hat{\pi}_h^k(\cdot|s), G_h^k(s,\cdot)\rangle.$$

*Moreover, the estimation error between $Q_h^{\pi^b}$, $V_h^{\pi^b}(s)$ and $\underset{\sim}{Q}_h^{k,b}$, $\underset{\sim}{V}_h^{k,b}$ can be bounded as*

$$Q_h^{\pi^b}(s,a) - \underset{\sim}{Q}_h^{k,b}(s,a) \le G_h^{k,b}(s,a),$$

$$V_h^{\pi^b}(s) - \underset{\sim}{V}_h^{k,b}(s) \le \langle \pi_h^b(\cdot|s), G_h^{k,b}(s,\cdot)\rangle.$$

**Lemma D.5** (Upper bound $\pi_1^k G_1^k$ and $\pi_1^{k,b} G_1^{k,b}$)**.** *Recall the functions of $G^k$ and $G^{k,b}$ in Equations (42) and (43). We have*

$$\pi_1^k G_1^k(s_1) \le 24e^{13}\sum_{h=1}^H\sum_{s,a} d_h^k(s,a)\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^k)(s,a)\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)}$$
$$+ 336e^{13}H^2\sum_{h=1}^H\sum_{s,a} d_h^k(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right),$$

*and*

$$\pi_1^{k,b} G_1^{k,b}(s_1) \le 24e^{13}\sum_{h=1}^H\sum_{s,a} d_h^b(s,a)\sqrt{\mathrm{Var}_{P_h}(V_{h+1}^{k,b})(s,a)\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)}$$
$$+ 336e^{13}H^2\sum_{h=1}^H\sum_{s,a} d_h^b(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right),$$

*where $d_h^k$ and $d_h^b$ are the occupancy measures under policy $\bar{\pi}^k$ and $\pi^b$, respectively.*

The proofs of the above three lemmas follow the same approaches as those for StepMix, and thus are omitted.

Besides, although the construction of the mixture policy under EpsMix is different from that under StepMix, the linearlity of the occupancy measure and the corresponding value function is preserved under EpsMix.

**Lemma D.6.** *Let $\pi = \rho\pi^1 \oplus (1-\rho)\pi^2$, and $d_h^1(s,a)$ and $d_h^2(s,a)$ be the occupancy measures under $\pi^1$ and $\pi^2$, respectively. Then, the following equality holds:*

$$d_h^\pi(s,a) = \rho d_h^1(s,a) + (1-\rho) d_h^2(s,a).$$

*Recall that the occupancy measure under a policy $\pi$ is defined as $d_h^\pi(s,a) = \mathbb{E}_\pi[\mathbb{1}\{s_h = s, a_h = a\}]$.*

*Proof.* Let $B_\rho$ be an independent Bernoulli random variable with mean $\rho$, and let $\pi$ be $\pi^1$ if $B_\rho = 1$ and be $\pi^2$ otherwise. Then,

$$\begin{aligned}
d_h^\pi(s,a) =& \mathbb{E}_\pi[\mathbb{1}\{s_h = s, a_h = a\}] \\
=& \mathbb{E}_\pi[\mathbb{1}\{s_h = s, a_h = a\} | B_\rho = 1]\mathbb{P}[B_\rho = 1] + \mathbb{E}_\pi[\mathbb{1}\{s_h = s, a_h = a\} | B_\rho = 0]\mathbb{P}[B_\rho = 0] \\
=& \mathbb{E}_{\pi^1}[\mathbb{1}\{s_h = s, a_h = a\}] \cdot \rho + \mathbb{E}_{\pi^2}[\mathbb{1}\{s_h = s, a_h = a\}](1-\rho) \\
=& \rho d_h^1(s,a) + (1-\rho) d_h^2(s,a).
\end{aligned}$$

□

**Proposition D.7.** *Under the same condition as in Lemma D.6, the following equality holds:*

$$V_1^\pi = \rho V_1^{\pi^1} + (1-\rho) V_1^{\pi^2}.$$

Based on the linearity shown in Lemma D.6, we obtain a result similar to that in Lemma C.8 for episodic mixture policies.

**Lemma D.8.** *For an episodic mixture policy $\pi^k$ mixed from two policies $\bar{\pi}^k$ and $\pi^b$, defined as $\pi^k = (1-\rho)\pi^b \oplus \rho\bar{\pi}^k$, the following bound holds:*

$$\begin{aligned}
&(1-\rho)\pi_1^b G_1^{k,b}(s_1) + \rho\bar{\pi}_1^k G_1^k(s_1) \\
&\leq 24e^{13} H \sqrt{\sum_{h=1}^H \sum_{s,a} d_h^k(s,a) \left( \frac{\beta(\bar{n}_h^k(s,a), \delta')}{\bar{n}_h^k(s,a) \vee 1} \right)} + 336e^{13} H^2 \sum_{h=1}^H \sum_{s,a} d_h^k(s,a) \left( \frac{\beta(\bar{n}_h^k(s,a), \delta')}{\bar{n}_h^k(s,a) \vee 1} \right),
\end{aligned} \tag{44}$$

*where $d_h^k(s,a)$ is the occupancy measure under policy $\pi^k$.*

Lemma D.8 can be proved following a similar approach as in the proof of Lemma C.8.

Now we establish the EpsMix version of the finite non-optimistic policy lemma.

**Lemma D.9.** *Define $\mathcal{N} = \{k | k \in [K], \pi^k \neq \bar{\pi}^k\}$. Then, the cardinality of $\mathcal{N}$ in the EpsMix algorithm can be bounded as*

$$|\mathcal{N}| \leq \left( \frac{4608e^{26}}{\kappa^2} + \frac{2688e^{13}S}{\kappa} \right) H^3 SA \log^2(K+1) = \tilde{O}\left( \left( \frac{1}{\kappa^2} + \frac{S}{\kappa} \right) H^3 SA \right),$$

*where $\kappa = V_1^{\pi^b} - \gamma$.*

*Proof.* If $\pi^k \neq \hat{\pi}^k$, we must have $\underset{\sim}{V}_1^k < \gamma = V_1^{\pi^b} - \kappa$. There are two possible cases for $\underset{\sim}{V}_1^{k,b}$. Case 1: $\underset{\sim}{V}_1^{k,b} < \gamma = V_1^{\pi^b} - \kappa$. For this case, the algorithm will choose $\pi^k = \pi^b$. Thus, $V_1^{\pi^b} - \underset{\sim}{V}_1^{k,b} > \kappa$. Case 2: $\underset{\sim}{V}_1^{k,b} \geq \gamma$. For this case, the algorithm will choose $\pi^k = \rho\bar{\pi}^k \oplus (1-\rho)\pi^b$. The design of $\rho$ ensures that $\rho\underset{\sim}{V}_1^k + (1-\rho)\underset{\sim}{V}_1^{k,b} = \gamma$. Therefore, $V_1^{\pi^b} - (\rho\underset{\sim}{V}_1^k + (1-\rho)\underset{\sim}{V}_1^{k,b}) = \kappa$. We note that $\pi^k = \pi^b$ can also be viewed as $1 \cdot \pi^b \oplus 0 \cdot \bar{\pi}^k$. Thus, for any $k \in \mathcal{N}$, we have

$$V_1^{\pi^b} - (\rho\underset{\sim}{V}_1^k + (1-\rho)\underset{\sim}{V}_1^{k,b}) \geq \kappa.$$

Furthermore, due to the optimality of $\pi^\star$, we have $V_1^{\pi^b} \leq \rho V_1^\star + (1-\rho)V_1^{\pi^b}$. Thus,

$$|\mathcal{N}|\kappa \leq \sum_{k \in \mathcal{N}} \left( V_1^{\pi^b} - (\rho\underset{\sim}{V}_1^k + (1-\rho)\underset{\sim}{V}_1^{k,b}) \right)$$

$$
\begin{aligned}
&\leq \sum_{k\in\mathcal{N}} \left( (\rho V_1^\star + (1-\rho) V_1^{\pi^b}) - (\rho \underset{\sim}{V}_1^k + (1-\rho) \underset{\sim}{V}_1^{k,b}) \right) \\
&= \sum_{k\in\mathcal{N}} \left( \rho (V_1^\star - \underset{\sim}{V}_1^k) + (1-\rho)(V_1^{\pi^b} - \underset{\sim}{V}_1^{k,b}) \right) \\
&\overset{(a)}{\leq} \sum_{k\in\mathcal{N}} \left( \rho \bar{\pi}_1^k G_1^k + (1-\rho) \pi_1^b G_1^{k,b} \right),
\end{aligned}
$$

where inequality $(a)$ is based on Lemma D.4.

Then, leveraging Lemma D.8, we have

$$
\begin{aligned}
|\mathcal{N}|\kappa &\leq 24e^{13} H \sum_{k\in\mathcal{N}} \sqrt{\sum_{h=1}^{H} \sum_{s,a} d_h^k(s,a) \left( \frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1} \right)} + 336 e^{13} H^2 \sum_{k\in\mathcal{N}} \sum_{h=1}^{H} \sum_{s,a} d_h^k(s,a) \left( \frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1} \right) \\
&\overset{(b)}{\leq} 24e^{13} H \sqrt{|\mathcal{N}|} \sqrt{\sum_{k\in\mathcal{N}} \sum_{h=1}^{H} \sum_{s,a} d_h^k(s,a) \left( \frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1} \right)} + 336 e^{13} H^2 \sum_{k\in\mathcal{N}} \sum_{h=1}^{H} \sum_{s,a} d_h^k(s,a) \left( \frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1} \right),
\end{aligned}
$$

where inequality $(b)$ follows from the Cauchy's inequality and $\delta' = \delta/4$.

Similar to Equation (36) and Equation (37) in Lemma C.9, we have

$$
\sum_{k\in\mathcal{N}} \sum_{h=1}^{H} \sum_{s,a} d_h^k(s,a) \left( \frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1} \right) \leq 4HSA\beta^\star(K,\delta') \log(|\mathcal{N}|+1), \tag{45}
$$

and

$$
\sum_{k\in\mathcal{N}} \sum_{h=1}^{H} \sum_{s,a} d_h^k(s,a) \left( \frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1} \right) \leq 4HSA\beta(K,\delta') \log(|\mathcal{N}|+1). \tag{46}
$$

Therefore, we have

$$
|\mathcal{N}|\kappa \leq 48e^{13} \sqrt{|\mathcal{N}|} \sqrt{H^3 SA \beta^\star(K,\delta') \log(|\mathcal{N}|+1)} + 1344 e^{13} H^3 SA \beta(K,\delta') \log(|\mathcal{N}|+1).
$$

By rearranging terms, we conclude that $|\mathcal{N}| \leq (\frac{4608e^{26}}{\kappa^2} + \frac{2688e^{13}S}{\kappa}) H^3 SA \log^2(K+1) = \tilde{O}((\frac{1}{\kappa^2} + \frac{S}{\kappa}) H^3 SA)$. □

Finally, we are ready to prove the regret upper bound of EpsMix.

*Proof of Theorem D.1.* First, we have

$$
\begin{aligned}
\sum_{k=1}^{K} \left( V_1^\star - V_1^{\pi^k} \right) &= \sum_{k\notin\mathcal{N}} \left( V_1^\star - V_1^{\pi^k} \right) + \sum_{k\in\mathcal{N}} \left( V_1^\star - V_1^{\pi^b} \right) + \sum_{k\in\mathcal{N}} \left( V_1^{\pi^b} - V_1^{\pi^k} \right) \\
&\leq \sum_{k\notin\mathcal{N}} (V_1^\star - \underset{\sim}{V}_1^k) + \sum_{k\in\mathcal{N}} \left( \rho^k (V_1^\star - \underset{\sim}{V}_1^k) + (1-\rho^k)(V_1^{\pi^b} - \underset{\sim}{V}_1^{k,b}) \right) + |\mathcal{N}|\Delta_0 \\
&= \sum_{k=1}^{K} \left( \rho^k (V_1^\star - \underset{\sim}{V}_1^k) + (1-\rho^k)(V_1^{\pi^b} - \underset{\sim}{V}_1^{k,b}) \right) + |\mathcal{N}|\Delta_0 \\
&\overset{(a)}{\leq} \sum_{k=1}^{K} \left( \rho^k \bar{\pi}_1^k G_1^k + (1-\rho^k) \pi_1^b G_1^{k,b} \right) + |\mathcal{N}|\Delta_0,
\end{aligned}
$$

where inequality $(a)$ follows from Lemma D.4.

Then, we use Lemma D.8 and the result of Lemma D.9 to bound the regret as follows:

$$
\begin{aligned}
&\sum_{k=1}^{K}\left(V_1^\star - V_1^{\pi^k}\right) \\
&\leq \sum_{k=1}^{K}\left(24e^{13}H\sqrt{\sum_{h=1}^{H}\sum_{s,a} d_h^k(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)} + 336e^{13}H^2\sum_{h=1}^{H}\sum_{s,a} d_h^k(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)\right) \\
&\quad + \left(\frac{4608e^{26}}{\kappa^2} + \frac{2688e^{13}}{\kappa}\right)H^3SA\log(K+1)\Delta_0 \\
&\leq 24e^{13}H\sqrt{K}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a} d_h^k(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)} + 336e^{13}H^2\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a} d_h^k(s,a)\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right) \\
&\quad + \left(\frac{4608e^{26}}{\kappa^2} + \frac{2688e^{13}S}{\kappa}\right)H^3SA\log(K+1)\Delta_0,
\end{aligned}
$$

where the last inequality is due to the Cauchy's inequality. Then we use the bound of summation in Lemma D.9 and plug Equations (45) and (46) into the above inequality, to conclude that

$$
\begin{aligned}
\mathrm{Reg}(K) &= \sum_{k=1}^{K}\left(V_1^\star - V_1^{\pi^k}\right) \\
&\leq 48e^{13}\sqrt{H^3SAK\beta^\star(K,\delta')\log(K+1)} + 1344e^{13}H^3SA\beta(K,\delta')\log(K+1) \\
&\quad + \left(\frac{4608e^{26}}{\kappa^2} + \frac{2688e^{13}S}{\kappa}\right)H^3SA\log^2(K+1)\Delta_0 \\
&= \tilde{O}\left(\sqrt{H^3SAK} + H^3S^2A + H^3SA\Delta_0\left(\frac{1}{\kappa^2} + \frac{S}{\kappa}\right)\right).
\end{aligned}
$$

□

## E. From Baseline Policy to Offline Dataset

### E.1. Offline Algorithm

The offline VI-LCB algorithm is detailed in Algorithm 5.

**Algorithm 5** Offline VI-LCB (Algorithm 3 in Xie et al. (2021))

**Require:** Dataset $D = \{(s_h^{(i)}, a_h^{(i)}, r_h^{(i)}, s_{h+1}^{(i)})_{h=1}^H\}_{i=1}^n$ collected using an unknown baseline policy $\mu$
  Randomly divide $D$ into $H$ sets $\{D_h\}_{h=1}^H$ such that $|D_h| = n/H$.
  Estimation $\hat{P}_h(s'|s,a)$ and $b_h(s,a)$ using $D_h$.
  Set $\hat{V}_{H+1}(s,a) = 0, \forall s, a$.
  **for** $h = H$ to 1 **do**
    $\hat{Q}_h(s,a) = \max(0, r_h(s,a) + \hat{P}_h\hat{V}_{h+1}(s,a) - b_h(s,a)), \forall s, a$.
    Let $\hat{\pi}_h(s) = \arg\max_a \hat{Q}_h(s,a), \forall s$.
    $\hat{V}_h(s) = \hat{Q}_h(s, \hat{\pi}_h(s)), \forall s$.
  **end for**
  Return $\hat{\pi}$.

### E.2. Theoretical Analysis

In Algorithm 5, $\hat{P}_h(s'|s,a)$ and $b_h(s,a)$ are defined as:

$$
\hat{P}_h(s'|s,a) = \frac{n_h(s,a,s')}{1 \vee n_h(s,a)}, \quad b_h(s,a) = c\sqrt{\frac{H^2\iota}{n_h(s,a)\vee 1}},
$$

where $\iota = \log(HSA/\delta)$, $n_h(s,a) = \sum_{s_h,a_h \in D_h} \mathbb{1}\{s_h = s, a_h = a\}$ is the count of visitations of state-action pair $(s,a)$ at step $h$, and $n_h(s,a,s') = \sum_{s_h,a_h,s_{h+1} \in D_h} \mathbb{1}\{s_h = s, a_h = a, s_{h+1} = s'\}$ is the count of visiting state-action pair $(s,a)$ at step $h$ while having state $s'$ as the next state. Both counts are only for samples in dataset $D_h$.

Following the approach in Xie et al. (2021), we first define the good events as follows.

**Lemma E.1** (Lemma B.1 in Xie et al. (2021))**.** *With probability at least $1-\delta$, there exists a finite constant $c$ such that the following good events hold:*

$$\forall h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}, \ \ |(P_h - \hat{P}_h)\hat{V}_{h+1}(s,a)| \le c\sqrt{\frac{H^2\iota}{n_h(s,a)}} = b_h(s,a), \quad \frac{1}{n_h(s,a)} \le c\frac{H\iota}{nd^\mu(s,a)},$$

*where $\iota = \log(HSA/\delta)$ and $d^\mu(s,a)$ is the occupancy measure under the behavior policy $\mu$.*

Under the good events, $\hat{Q}_h(s,a)$ can be proved to be the lower confidence bound of $Q_h^{\hat{\pi}}(s,a)$ as shown in the following lemma.

**Lemma E.2** (Lemma B.2 in Xie et al. (2021))**.** *Let $\hat{Q}_h(s,a) = \max(0, r_h(s,a) + \hat{P}_h\hat{V}_{h+1}(s,a) - b_h(s,a))$. Then, under the good events defined in Lemma E.1, we have*

$$\hat{Q}_h(s,a) \le Q_h^{\hat{\pi}}(s,a).$$

The above two lemmas are the same as Xie et al. (2021), and thus we omit the proofs.

We first bound the value difference $V_1^\mu - V_1^{\hat{\pi}}$ with the offline estimation bonus $b_h^k(s,a)$ in the following lemma.

**Lemma E.3.** *Suppose there are $n$ trajectories collected under the behavior policy $\mu$. Then, under the good events defined in Lemma E.1, the extracted policy $\hat{\pi}$ satisfies*

$$V_1^\mu - V_1^{\hat{\pi}} \le 2\sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\mu(s,a) b_h(s,a),$$

*where $b_h(s,a) = c\sqrt{\frac{H^2\iota}{n_h(s,a)\vee 1}}$.*

*Proof.* We directly calculate the suboptimality gap as follows

$$\begin{aligned}
V_h^\mu(s) - V_h^{\hat{\pi}}(s) =& V_h^\mu(s) - \max_a Q_h^{\hat{\pi}}(s,a) \\
\le& \mathbb{E}_{a\sim\mu_h(\cdot|s)}[Q_h^\mu(s,a) - Q_h^{\hat{\pi}}(s,a)] \\
\le& \mathbb{E}_{a\sim\mu_h(\cdot|s)}[b_h(s,a) + P_h V_{h+1}^\mu(s,a) - \hat{P}_h\hat{V}_{h+1}(s,a)] \\
=& \mathbb{E}_{a\sim\mu_h(\cdot|s)}[b_h(s,a) + P_h(V_{h+1}^\mu - \hat{V}_{h+1})(s,a) + (P_h - \hat{P}_h)\hat{V}_{h+1}(s,a)] \\
\le& 2\mathbb{E}_{a\sim\mu_h(\cdot|s)}[b_h(s,a)] + \mathbb{E}_{a\sim\mu_h(\cdot|s), s'\sim P(\cdot|s,a)}[V_{h+1}^\mu(s') - V_{h+1}^{\hat{\pi}}(s')].
\end{aligned}$$

Recursively unfolding the above inequality from $h=1$, we have:

$$V_1^\mu - V_1^{\hat{\pi}} \le 2\sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\mu(s,a) b_h(s,a).$$

□

The following theorem establishes an upper bound for the gap between the learned policy $\hat{\pi}$ and the behavior policy $\mu$.

**Theorem E.4** (Adapted from Theorem 1 in Xie et al. (2021))**.** *Suppose $n$ trajectories are collected in the offline dataset collected under policy $\mu$. Then, with probability at least $1-\delta$, the output policy $\hat{\pi}$ of the offline Algorithm VI-LCB satisfies*

$$V_1^\mu - V_1^{\hat{\pi}} \le 2c\iota\sqrt{\frac{H^5SA}{n}},$$

*where $\iota = \log(HSA/\delta)$.*

*Proof.* Under the good events defined in Lemma E.1, we have

$$
\begin{aligned}
V_1^\mu - V_1^{\hat{\pi}} &\overset{(a)}{\leq} 2\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\mu(s,a) b_h(s,a) \\
&\leq 2c\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\mu(s,a)\sqrt{\frac{H^2\iota}{n_h(s,a)\vee 1}} \\
&\overset{(b)}{\leq} 2c\sqrt{H^2\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\mu(s,a)\sqrt{\frac{H\iota}{nd_h^\mu(s,a)}} \\
&\leq 2c\iota\sqrt{\frac{H^3}{n}}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{d_h^\mu(s,a)} \\
&\overset{(c)}{\leq} 2c\iota\sqrt{\frac{H^3}{n}}\sqrt{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} 1}\sqrt{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^\mu(s,a)} \\
&= 2c\iota\sqrt{\frac{H^5SA}{n}},
\end{aligned}
$$

where $(a)$ is from Lemma E.3, $(b)$ follows from the definition of good events in Lemma E.1, and $(c)$ is based on the Cauchy's inequality. □

Based on the above theorem, we have the following corollary regarding the sample complexity.

**Corollary E.5.** *With probability at least $1-\delta/2$, if $n \geq \frac{16c^2\iota'^2H^5SA}{(V_1^\mu-\gamma)^2}$ and $V_1^\mu > \gamma$, the output $\hat{\pi}$ of offline VI-LCB satisfies $V_1^{\hat{\pi}} \geq (V_1^\mu+\gamma)/2$, where $\iota' = \log(2HSA/\delta)$.*

*Proof.* To ensure $V_1^{\hat{\pi}} > (V_1^\mu+\gamma)/2$, we need to establish $V_1^\mu - V_1^{\hat{\pi}} < (V_1^\mu-\gamma)/2$. We note that if

$$
2c\iota\sqrt{\frac{H^5SA}{n}} < (V_1^\mu-\gamma)/2, \tag{47}
$$

then $V_1^\mu - V_1^{\hat{\pi}} < (V_1^\mu-\gamma)/2$. Rearranging the terms in Equation (47) leads to

$$
n > \frac{16c^2\iota'^2H^5SA}{(V_1^\mu-\gamma)^2},
$$

which completes the proof. □

Combining Corollary E.5 and Theorem 4.2, we can prove the following theorem.

**Theorem E.6.** *Assume that there are at least $n \geq \frac{16c^2\iota'^2H^5SA}{(V_1^\mu-\gamma)^2}$ offline trajectories collected under a safe behavior policy $\mu$. If we let Algorithm 5 run on the offline dataset and pass the output $\hat{\pi}$ to Algorithm 1 as the baseline $\pi^b$, then, with probability at least $1-\delta$, StepMix does not violates the constraint in Equation (2) and achieves a regret that scales in*

$$
\tilde{O}\left(\sqrt{H^3SAK} + H^3S^2A + H^3SA\bar{\Delta}_0\left(\frac{1}{\bar{\kappa}^2}+\frac{S}{\bar{\kappa}}\right)\right),
$$

*where $\bar{\kappa} = (V_1^\mu-\gamma)/2 > 0$ and $\bar{\Delta}_0 = V_1^\star - V_1^\mu + \bar{\kappa}$.*

*Proof.* Corollary E.5 states that with $n \geq \frac{16c^2\iota'^2H^5SA}{(V_1^\mu-\gamma)^2}$ offline trajectories, $\mathbb{P}[V_1^{\hat{\pi}} > (V_1^\mu+\gamma)/2] \geq 1-\delta/2$. Denote $A$ the event that $\hat{\pi}$ satisfies $V^{\hat{\pi}} > (V_1^\mu+\gamma)/2$. Then, we have $\mathbb{P}[A] \geq 1-\delta/2$.

Theorem 4.2 states that if $V^{\hat{\pi}} > (V_1^{\mu} + \gamma)/2 > \gamma$, with probability $1 - \delta/2$, StepMix does not violate the constraint and achieves a regret at most

$$\begin{aligned} &48e^{13}\sqrt{H^3SAK\beta^{\star}(K,\delta')\log(K+1)} + 1344e^{13}H^3SA\beta(K,\delta')\log(K+1) \\ &\quad + \left(\frac{4608e^{26}}{\bar{\kappa}^2} + \frac{2688e^{13}S}{\bar{\kappa}}\right)H^3SA\log^2(K+1)\Delta_0 \\ &= \tilde{O}\left(\sqrt{H^3SAK} + H^3S^2A + H^3SA\bar{\Delta}_0\left(\frac{1}{\bar{\kappa}^2} + \frac{S}{\bar{\kappa}}\right)\right), \end{aligned}$$

where $\delta' = \frac{\delta}{6(H+1)}$, $V_1^{\hat{\pi}} - \gamma \geq \bar{\kappa} = (V_1^{\mu} - \gamma)/2$, and $\bar{\Delta}_0 = V_1^{\star} - V_1^{\mu} + \bar{\kappa}$.

Since we use $\hat{\pi}$ as baseline, by letting $B$ denote the event that StepMix achieves the regret in Theorem E.6 and does not violate the constraint, we have $\mathbb{P}[B|A] \geq 1 - \delta/2$. Because $\mathbb{P}[B] = \mathbb{P}[B|A]\mathbb{P}[A] = (1-\delta/2)(1-\delta/2) \geq 1 - \delta$, we have that, with overall probability at least $1 - \delta$, when StepMix uses the output of offline UCB-VI as the baseline policy, it achieves a regret that is at most

$$\tilde{O}\left(\sqrt{H^3SAK} + H^3S^2A + H^3SA\bar{\Delta}_0\left(\frac{1}{\bar{\kappa}^2} + \frac{S}{\bar{\kappa}}\right)\right)$$

without violating the constraint. □

We can also combine Corollary E.5 and Theorem D.1 to prove the following theorem for EpsMix, which is similar to Theorem E.6 for StepMix.

**Theorem E.7.** *Assume that there are at least $n \geq \frac{16c^2\iota'^2H^5SA}{(V_1^{\mu}-\gamma)^2}$ trajectories collected under a safe behavior policy $\mu$. If we let Algorithm 5 run on the offline dataset and pass the output $\hat{\pi}$ to Algorithm 4 as the baseline $\pi^b$, then, with probability at least $1 - \delta$, EpsMix does not violate the constraint in Equation (2) and achieves a regret that scales in*

$$\tilde{O}\left(\sqrt{H^3SAK} + H^3S^2A + H^3SA\bar{\Delta}_0\left(\frac{1}{\bar{\kappa}^2} + \frac{S}{\bar{\kappa}}\right)\right),$$

*where $\bar{\kappa} = (V_1^{\mu} - \gamma)/2 > 0$ and $\bar{\Delta}_0 = V_1^{\star} - V_1^{\mu} + \bar{\kappa}$.*

The proof is similar to that of Theorem E.6 and is thus omitted.

## F. Technical Lemmas

In this section, we list several technical lemmas that are used in the main proof.

**Lemma F.1** (Lemma 11 in Ménard et al. (2021))**.** *Let $p$ and $q$ be two probability distributions supported by the state set $\mathcal{S}$, and $f$ be a function on $\mathcal{S}$. If KL$(p,q) \leq \alpha$, $0 \leq f(s) \leq b, \forall s \in \mathcal{S}$, we have*

$$\mathrm{Var}_q(f) \leq 2\mathrm{Var}_p(f) + 4b^2\alpha,$$

$$\mathrm{Var}_p(f) \leq 2\mathrm{Var}_q(f) + 4b^2\alpha.$$

**Lemma F.2** (Lemma 12 in Ménard et al. (2021))**.** *Let $p$ and $q$ be two probability distributions supported by the state set $\mathcal{S}$, and $f, g$ be functions on $\mathcal{S}$. If $0 \leq g(s), f(s) \leq b, \forall s \in \mathcal{S}$, we have*

$$\mathrm{Var}_p(f) \leq 2\mathrm{Var}_p(g) + 2b\mathbb{E}_p[|f-g|],$$

$$\mathrm{Var}_q(f) \leq \mathrm{Var}_p(f) + 3b^2\left\|p-q\right\|_1.$$

**Lemma F.3** (Lemma 10 in Ménard et al. (2021))**.** *Let $p$ and $q$ be two probability distributions supported by the state set $\mathcal{S}$, and $f$ be a function on $\mathcal{S}$. If KL$(p,q) \leq \alpha$ and $0 \leq f \leq b$, we have*

$$|\mathbb{E}_p[f] - \mathbb{E}_q[f]| \leq \sqrt{2\mathrm{Var}_q(f)\alpha} + \frac{2}{3}b\alpha.$$

**Lemma F.4** (Lemma 6 in Huang et al. (2022)). *Given transition kernel $P_h$, policy $\pi$ and reward $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$, we have:*

$$\sum_{h=1}^{H} \sum_{s,a} d_h^\pi(s,a) \mathrm{Var}_{P_h}(V_{h+1}^\pi)(s,a) = \mathbb{E}_{\pi,P}\left[\left(\sum_{h=1}^{H} r_h(s_h,a_h) - V_1^\pi(s_1)\right)^2\right] \le H,$$

*where $d_h^\pi$ is the occupancy measure under policy $\pi$ and $V_h^\pi$ is the value function.*

**Lemma F.5.** *Under event $\mathcal{E}^{cnt}$ and using the same notations defined in Appendix B, we have*

$$\left(\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \wedge 1\right) \le 4\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1}\right).$$

*Similarly, for $\beta^\star$, the following inequality holds:*

$$\left(\frac{\beta^\star(n_h^k(s,a),\delta')}{n_h^k(s,a)} \wedge 1\right) \le 4\left(\frac{\beta^\star(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a) \vee 1}\right).$$

*Proof.* Event $\mathcal{E}^{\text{cnt}}$ means that

$$n_h^k(s,a) \ge \frac{1}{2}\bar{n}_h^k(s,a) - \beta^{\text{cnt}}(\delta').$$

If $\beta^{\text{cnt}}(\delta') \le \frac{1}{4}\bar{n}_h^k(s,a)$, we directly have the result $n_h^k(s,a) \ge \frac{1}{4}\bar{n}_h^k(s,a)$, which proves $\left(\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \wedge 1\right) \le 4\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)$. On the other hand, if $\beta^{\text{cnt}}(\delta') \ge \frac{1}{4}\bar{n}_h^k(s,a)$, based on the fact that $\beta(n_h^k(s,a),\delta') \ge \beta^{\text{cnt}}(\delta') > 1$, we have that $\left(\frac{\beta(n_h^k(s,a),\delta')}{n_h^k(s,a)} \wedge 1\right) \le 1 \le 4\left(\frac{\beta(\bar{n}_h^k(s,a),\delta')}{\bar{n}_h^k(s,a)\vee 1}\right)$. The same arguments can also be applied to the inequality with $\beta^\star$. □

**Lemma F.6.** *For any state-action pair $(s,a)$, step $h$ and a subset of episodes $\mathcal{N} \subseteq [K]$, we have*

$$\sum_{k\in\mathcal{N}} \frac{d_h^k(s,a)}{\bar{n}_h^k(s,a) \vee 1} \le 4\log(|\mathcal{N}| + 1),$$

*where $d_h(s,a)$ is the occupancy measure, $\bar{n}_h^k(s,a)$ is the expected visitation count.*

*Proof.* We have

$$\begin{aligned}
\sum_{k\in\mathcal{N}} \frac{d_h^k(s,a)}{\bar{n}_h^k(s,a) \vee 1} &= \sum_{k\in\mathcal{N}} \frac{d_h^k(s,a)}{(\sum_{t=1}^{h-1} d_t^k(s,a)) \vee 1} \\
&\le \sum_{k\in\mathcal{N}} \frac{d_h^k(s,a)}{(\sum_{t\in\mathcal{N},t<k} d_t^k(s,a)) \vee 1} \\
&\le \sum_{k\in\mathcal{N}} \frac{4d_h^k(s,a)}{2\sum_{t\in\mathcal{N},t<k} d_t^k(s,a) + 2} \\
&\le 4\sum_{k\in\mathcal{N}} \frac{d_h^k(s,a)}{\sum_{t\in\mathcal{N},t<k} d_t^k(s,a) + d_h^k(s,a) + 1} \\
&\le 4\sum_{k\in\mathcal{N}} \frac{d_h^k(s,a)}{\sum_{t\in\mathcal{N},t\le k} d_t^k(s,a) + 1}.
\end{aligned}$$

Then, by noting $f(k) = \sum_{t\in\mathcal{N},t\le k} d_t^k(s,a)$ and $k' = \max_t\{t \in \mathcal{N} \cup \{0\} | t < k\}$, we have

$$4\sum_{k\in\mathcal{N}} \frac{d_h^k(s,a)}{\sum_{t\in\mathcal{N},t\le k} d_t^k(s,a) + 1} \le 4\sum_{k\in\mathcal{N}} \frac{f(k) - f(k')}{f(k) + 1}$$

$$
\begin{aligned}
&\leq 4 \sum_{k \in \mathcal{N}} \int_{x=f(k')}^{f(k)} \frac{dx}{x+1} \\
&\leq 4 \int_{x=1}^{|\mathcal{N}|+1} \frac{1}{x} dx \\
&= 4 \log(|\mathcal{N}|+1).
\end{aligned}
$$

□

**Theorem F.7** (Proposition 1 in Jonsson et al. (2020))**.** *For a categorical distribution with probability distribution $p \in \Sigma_m$, denoting $\hat{P}_n$ as a frequency estimation of $p$, we have*

$$
\mathbb{P}\Big(\exists n \in \mathbb{N}^\star, nKL(\hat{P}_n, p) > \log(1/\delta) + (m-1)\log(e(1+n/(m-1))\Big) \leq \delta.
$$

**Theorem F.8** (Lemma F.4 in Dann et al. (2017))**.** *Let $\{\mathcal{F}_t\}_{t=1}^n$ be a filtration, $\{X_t\}_{t=1}^n$ be a series of Bernoulli random variables with $\mathbb{P}[X_t = 1|\mathcal{F}_{t-1}] = p_t$, where $p_t$ is $\mathcal{F}_{t-1}$-measurable. Then*

$$
\forall \delta > 0, \mathbb{P}\left(\exists n : \sum_{t=1}^n X_t < \sum_{t=1}^n p_t/2 - \log(1/\delta)\right) \leq \delta.
$$

**Theorem F.9** (Theorem 5 in Ménard et al. (2021), Lemma 3 in (Domingues et al., 2021b))**.** *Suppose $(Y_t)_{t\in\mathbb{N}}$ and $(w_t)_{t\in\mathbb{N}}$ are two sequences from filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$, subject to $w_t \in [0,1]$, $|Y_t| \leq b$ and $\mathbb{E}[Y_t|\mathcal{F}_t] = 0$. Define*

$$
\mathcal{S}_t = \sum_{s=1}^t w_s Y_s, \quad \mathcal{V}_t = \sum_{s=1}^t w_s^2 \mathbb{E}[Y_s^2|\mathcal{F}_s],
$$

*then, for any $\delta \in (0,1)$, we have*

$$
\mathbb{P}\left(\exists t \geq 1, (\mathcal{V}_t/b^2 + 1) h\left(\frac{b|\mathcal{S}_t|}{\mathcal{V}_t + b^2}\right) \geq \log(1/\delta) + \log(4e(2t+1))\right) \leq \delta,
$$

*where $h(x) = (x+1)\log(x+1) - x$.*

*This result can be equivalently stated as: with probability at least $1-\delta$, the following inequality holds:*

$$
|\mathcal{S}_t| \leq \sqrt{2\mathcal{V}_t \log 4e(2t+1)/\delta} + 3b\log(4e(2t+1)/\delta).
$$