CONSERVATIVE EXPLORATION IN LINEAR MDPS UNDER EPISODE-WISE CONSTRAINTS

Anonymous authors
Paper under double-blind review

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 episodic linear Markov Decision Process (MDP) setting where the transition kernels and the reward functions are assumed to be linear. With the knowledge of an existing safe baseline policy, two algorithms based on Least-Squares Value Iteration (LSVI) (Bradtke & Barto, 1996; Osband et al., 2016), coined StepMix-LSVI and EpsMix-LSVI, are proposed to balance the exploitation and exploration while ensuring that the conservative constraint is never violated in each episode with high probability. Theoretical analysis shows that both algorithms achieve the same regret order as LSVI-UCB, their constraint-free counterpart from Jin et al. (2020), indicating that obeying the stringent episode-wise conservative constraint does not compromise the learning performance of these algorithms. The algorithm design and the 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.

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 guarantees 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. 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., 2020b; Yang et al., 2021b; 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 (see Appendix A for a detailed literature review) 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.

Besides the difficulties of incorporating the conservative constraints, practical RL applications may involve a large number of states, and *linear function approximation* (Jin et al., 2020; Yang & Wang,

Table 1: Comparison of Conservative Exploration Methods with Linear Function Approximation

Algorithm (Reference)	Regret	Constraint	Additional Assumption
LSVI-UCB (Jin et al., 2020)	$\tilde{O}\left(\sqrt{d^3H^4N}\right)$	-	-
SLUCB-QVI (Amani et al., 2021)	$\tilde{O}(\frac{1}{\kappa'}\sqrt{d^3H^4N})$	Step	Continuous feature space, known κ'
BudgetFirst / LCBCE (Yang et al., 2021b)	$\tilde{O}\left(\sqrt{d^3H^4N} + \frac{d^3H^4\Delta_0}{\kappa^2+\kappa\Delta_0}\right)$	Cumulative	Known / Unknown Δ_0
StepMix-LSVI (this work)	$\tilde{O}\left(\sqrt{d^3H^4N} + \frac{d^3H^4\Delta_0}{\kappa^2}\right)$	Episodic	-
EpsMix-LSVI (this work)	$\tilde{O}\left(\sqrt{d^3H^4N} + \frac{d^3H^4\Delta_0}{\kappa^2}\right)$	Episodic	-

d: dimension; H: horizon; N: number of episodes; Δ_0: suboptimality gap for the base policy;
κ: tolerable reward loss from base policy. κ': minimum gap between the costs of base actions and the constraint.

2020) is a simple yet effective approach to approximate either the value function or the policy. The introduction of linear function approximation raises a fundamental set of challenges involving the computational and statistical efficiency, especially given the need to manage the exploration/exploitation tradeoff. Incorporating linear function approximation also makes the conservative exploration problem much more difficult (Amani et al., 2021; Yang et al., 2021b).

In this paper, we focus on conservative exploration in a finite-horizon episodic MDP with linear function approximation. Unlike most of the prior works, we enforce a more strict conservative constraint that the expected reward of the RL policy cannot be much worse than that of a baseline policy *for every episode*. When such baseline policy is explicitly given, we propose to integrate various types of mixture policies into conservative exploration to cope with the more stringent per-episode constraint. We then extend the study of conservative exploration to the setting where the baseline policy is not given a priori but must be learned from an offline dataset. Our main contributions are summarized as follows.

- With a given baseline policy, we propose StepMix-LSVI, a new model-free learning algorithm based on the Least-Squares Value Iteration (LSVI) principle with a step mixture design embedded in each episode. StepMix-LSVI is built on a novel two-stage policy design that dynamically evolves with episodes, where we add an *evaluation* step to examine different concatenations of the baseline policy and the optimistic policy in terms of *potential* constraint violations. This evaluation relies on a careful integration of the LSVI principle and the lower confidence bound (LCB) to produce the desired balance of baseline and optimistic policies.
- We then develop EpsMix-LSVI which, instead of relying on step mixture policies as in StepMix-LSVI, adopts a randomization mechanism and constructs episodic mixture policies in each episode. *Episode-wise randomization* is critical in allowing EpsMix-LSVI to be less conservative than StepMix-LSVI without violating the conservative constraint.
- Regret analyses reveal that without any additional assumption, both algorithms achieve $\tilde{O}\left(\sqrt{d^3H^4N}\right)$ regret, which is of the same order as LSVI-UCB, their constraint-free counterpart (Jin et al., 2020), 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 Amani et al. (2021). Furthermore, the additive terms are comparable to that in Yang et al. (2021b), while our constraint is more stringent. A comparison of our work and these relevant papers is presented in Table 1.
- When the baseline policy is not given, we study an extension where we first learn an approximately safe baseline policy from an offline dataset and then use it as an input to the StepMix-LSVI or EpsMix-LSVI algorithm. We characterize the impact of safety uncertainty due to learning from an offline dataset on the safety and regret of conservative exploration, and prove that as long as the dataset is sufficiently large, similar regret and conservative guarantees to the case of explicitly using a provably safe baseline policy can be achieved. This is further validated in the experiments.

2 Problem Formulation

We consider an episodic MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r, x_1)$, 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 x_1 is a given initial state. We assume that $\mathcal{S}$ is a measurable space with possibly infinite number of elements and $\mathcal{A}$ is a finite set with cardinality A. Moreover, for each $h \in [H]$, $P_h(\cdot|x,a)$ denotes the transition kernel over the next states if action a is taken for state x at step $h \in [H]$, and $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the deterministic reward function at step h. Our result can be easily generalized to random reward functions. We consider the learning problem where $\mathcal{S}$ and $\mathcal{A}$ are known while P and r are unknown a priori.

A policy π 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 x_1. Then, at each step $h \in [H]$, the agent observes the state $x_h \in \mathcal{S}$, picks an action $a_h \in \mathcal{A}$, and receives a reward $r_h(x_h, a_h) \in [0,1]$. The MDP then evolves into a new state x_{h+1} that is drawn from the probability measure $P_h(\cdot|x_h, a_h)$. 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 π when starting from an arbitrary state at the h-th step until the end of the episode. Specifically, $\forall x \in \mathcal{S}, h \in [H]$,

$$V_h^\pi(x) := \mathbb{E}_\pi\bigg[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'})\bigg| x_h = x\bigg], \tag{1}$$

where we use $\mathbb{E}_\pi[\cdot]$ to denote the expectation over states and actions that are governed by π and P. Since the MDP begins with the same initial state x_1, to simplify the notation, we use V^π to denote $V_1^\pi(x_1)$ 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 π after taking action a at state x in step h, that is:

$$Q_h^\pi(x,a) := \mathbb{E}_\pi\bigg[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'})\bigg| x_h = x, a_h = a\bigg] = r_h(x,a) + [\mathbb{P}_h V_{h+1}^\pi](x,a),$$

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

Linear MDP. We assume the MDP $(\mathcal{S}, \mathcal{A}, H, P, r, x_1)$ is a linear MDP (Jin et al., 2020) with a (known) feature map ϕ, i.e., for any $h \in [H]$, there exist d unknown measures $\mu_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$, such that for any $(x,a) \in \mathcal{S} \times \mathcal{A}$, we have $P_h(x'|x,a) = \langle \phi(x,a)$ and $\mu_h(x')\rangle, r_h(x,a) = \langle \phi(x,a), \theta_h\rangle$. Without loss of generality, we assume $\|\phi(x,a)\| \leq 1$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$, and $\max\{\|\mu_h(\mathcal{S})\|, \|\theta_h\|\} \leq \sqrt{d}$ for all $h \in [H]$.

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., 2020b; Yang et al., 2021b). 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 π^b and the corresponding expected total reward obtained under π^b in an episode as V^{π^b}. Then, throughout the learning, we require that the expected total reward for each episode n is at least γ with high probability, where $\kappa := V^{\pi^b} - \gamma > 0$ characterizes how much risk the algorithm can take during the learning process. A policy π that achieves expected total reward at least γ is considered to be "safe", and we emphasize that our proposed algorithms do not require the knowledge of V^{π^b}. Let π^n be the policy adopted by the agent during episode $n \in [N]$. Mathematically, we formulate the conservative constraint as

$$\mathbb{P}\Big[V^{\pi^n} \geq \gamma, \forall n \in [N]\Big] \geq 1 - \delta, \text{ where } \delta \in (0,1). \tag{2}$$

Comparison with Previous Conservative Constraints. The conservative constraint in Equation (2) is more restrictive compared with Garcelon et al. (2020b); Yang et al. (2021b), 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. A comprehensive review of the related works and how our work is different from them are deferred to Appendix A due to the space limitation.

In addition, the per-episode conservative constraint in our paper 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^{π^n} and V^* serves as the expected regret or the suboptimality of the agent in the n-th episode. Thus, after playing for N episodes, the total expected regret is

$$\text{Reg}(N) := NV^* - \sum_{n=1}^{N} V^{\pi^n}. \tag{3}$$

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

3 The StepMix-LSVI Algorithm

In this section, we present a new model-free learning algorithm based on Least-Squares Value Iteration (Bradtke & Barto, 1996; Osband et al., 2016), coined StepMix-LSVI. A major component of StepMix-LSVI is to design a step mixture policy (Baram et al., 2021) in each episode. Before we present the StepMix-LSVI algorithm, we first introduce the definition of step mixture policies.

Definition 1 (Step Mixture Policy)**.** The step mixture policy of two Markov policies π^1 and π^2 with a parameter ρ, 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 x_h under the step mixture policy is $\rho\pi_h^1(a_h|x_h) + (1-\rho)\pi_h^2(a_h|x_h)$.

3.1 Algorithm Design

We elaborate the design of StepMix-LSVI, which is presented in Algorithm 1. The StepMix-LSVI algorithm proceeds in episodes. At the beginning of each episode n, it first constructs an *information matrix* Λ_h^n for each step h. This information matrix, together with the historical data, will be utilized in the LSVI-UCB subroutine to obtain an optimistic policy $\bar{\pi}^n$. The LSVI-UCB subroutine, shown in Algorithm 2 in Appendix C.1, is directly adopted from Jin et al. (2020). The essential idea of LSVI-UCB is to parameterize $Q_h^*(x, a)$ by the linear form $w_h^\top \phi(x, a)$, where w_h can be approximated by solving a least-square problem. An upper confidence bound (UCB) is then added to the estimated $Q_h^*(x, a)$ to balance exploration and exploitation during learning. When the conservative constraint is not imposed, LSVI-UCB is known to converge to the optimal policy with a regret in the order of $O(\sqrt{d^3H^4N})$. However, LSVI-UCB cannot guarantee that the total expected reward obtained in each episode is always above the desired threshold γ. In other words, the necessary exploration may temporarily hurt the total reward in certain episodes and lead to undesirable performance degradation.

To overcome this disadvantage of LSVI-UCB and avoid constraint violation during learning, StepMix-LSVI relies on a novel two-stage construction to produce the mixture policy π^n for episode n. Notably, this two-stage construction dynamically evolves with episodes. In the **first stage**, we construct a set of candidate policies by concatenating the baseline policy π^b with the optimistic policy $\bar{\pi}^n$. The purpose of this set of candidate policies is to evaluate potential violations of the

constraint, which directly affects the mixture probability in Definition 1. Specifically, we let the first $h_0 \in [0, H]$ steps of the candidate policy be the same as the baseline policy π^b, and the last $H - h_0$ steps be the same as the optimistic policy $\bar{\pi}^n$. We denote each of those $H + 1$ policies as $\tilde{\pi}^{n,h_0}$, and then estimate the performance of those policies by invoking the LCB-V subroutine (see Algorithm 3 in Appendix C.2).

The LCB-V subroutine also utilizes the LSVI principle to estimate the action-value function $Q_h^{\tilde{\pi}^{n,h_0}}(\cdot,\cdot)$ under policy $\tilde{\pi}^{n,h_0}$. Compared with LSVI-UCB, there are two major differences: First, instead of constructing an upper confidence bound for the action-value function, we obtain a lower confidence bound (LCB) by subtracting the bonus term $\beta\|\phi(\cdot,\cdot)\|_{(\Lambda_h^n)^{-1}}$, where we use $\|x\|_V$ to denote $\sqrt{x^\top V x}$. We expect that the true action-value function is above its LCB with high probability, similarly for the state-value function. Second, since LCB-V is a policy *evaluation* subroutine, it does not apply the $\max$ operator to the action-value function for the greedy action selection. Rather, in order to evaluate the state-value function, it needs to take expectation of the action-value function. These differences would also complicate the corresponding analysis.

Once the LCB of the expected total reward $\underline{V}^{\tilde{\pi}^{n,h_0}}$ is returned, the learner will compare it with the threshold γ. If it is above the threshold, it indicates that with high probability the concatenated policy $\tilde{\pi}^{n,h_0}$ will satisfy the conservative constraint. Then, the learner would stop evaluating the remaining concatenated policies. Let h_n be the smallest h_0 such that $\underline{V}^{\tilde{\pi}^{n,h_0}} \geq \gamma$. If none of the candidate policies achieves the threshold, we let $h_n = H + 1$.

Algorithm 1 The StepMix-LSVI Algorithm

Input: λ, β, γ, π^b
$\mathcal{D}_0 \leftarrow \emptyset$
for episode $n = 1, 2, \ldots, N$ **do**
  // `1st stage: policy evaluation`
  $\Lambda_h^n \leftarrow \sum_{\tau=1}^{n-1} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + \lambda I$
  $\bar{\pi}^n \leftarrow$ LSVI-UCB$(\mathcal{D}_{n-1}, \Lambda_h^n)$; $h_0 \leftarrow 0$
  while $h_0 \leq H$ **do**
    $\tilde{\pi}^{n,h_0} \leftarrow (\pi_1^b, \ldots, \pi_{h_0}^b, \bar{\pi}_{h_0+1}^n, \ldots, \bar{\pi}_H^n)$
    $\underline{V}^{\tilde{\pi}^{n,h_0}} \leftarrow$ LCB-V$(\tilde{\pi}^{n,h_0}, \mathcal{D}_{n-1}, \Lambda_h^n, \beta)$
    if $\underline{V}^{\tilde{\pi}^{n,h_0}} \geq \gamma$ **then**
      Break;
    end if
    $h_0 \leftarrow h_0 + 1$
  end while
  $h_n \leftarrow h_0$
  // `2nd stage: policy construction`
  if $h_n \in [1 : H]$ **then**

$$\rho_n = \frac{\underline{V}^{\tilde{\pi}^{n,h_n}} - \gamma}{\underline{V}^{\tilde{\pi}^{n,h_n}} - \underline{V}^{\tilde{\pi}^{n,h_n-1}}} \tag{4}$$

$$\pi^n = \rho_n \tilde{\pi}^{n,h_n-1} + (1 - \rho_n)\tilde{\pi}^{n,h_n} \tag{5}$$

  else if $h_n = 0$ **then**
    $\pi^n = \bar{\pi}^n$
  else
    $\pi^n = \pi^b$
  end if
  Play π^n and collect $\{(x_h^n, a_h^n, r_h(x_h^\tau, a_h^\tau))\}_{h=1}^H$
  $\mathcal{D}_n \leftarrow \mathcal{D}_{n-1} \cup \{(x_h^n, a_h^n, r_h(x_h^\tau, a_h^\tau))\}_{h=1}^H$
end for

Then, in the **second stage**, depending on the value of h_n, the learner constructs and executes the policy π^n for episode n as follows:

- If $h_n = 0$, LSVI-UCB is considered safe, and the learner executes $\bar{\pi}^n$.
- If $h_n \in [1 : H]$, it indicates that $\tilde{\pi}^{n,h_n}$ is safe but $\tilde{\pi}^{n,h_n-1}$ may be not. More importantly, they only differ in one step h_n. Then, the learner would construct a mixture of $\tilde{\pi}^{n,h_n}$ and $\tilde{\pi}^{n,h_n-1}$ according to Equation (4) and Equation (5), such that the expected total reward obtained under the mixture policy is guaranteed to be above the threshold.
- If $h_n = H + 1$, it indicates that the LCB of V^{π^b} is below the threshold, which occurs when the estimation is highly uncertain. The learner will then resort to π^b for conservative exploration.

Therefore, at each episode n, StepMix-LSVI finds a safe policy π^n chosen from either π^b, $\bar{\pi}^n$, or a step mixture policy in Equation (5) . Once the policy π^n is executed and a trajectory is collected, the learner moves on to the next episode.

3.2 Theoretical Analysis

The performance of StepMix-LSVI is formally stated in the following theorem.

Theorem 1. *There exist absolute constants c', c_β, c_1 and c_2 such that, for any $\delta \in (0, 1)$, if we choose $\lambda = c' d \log(dNH/\delta)$ and $\beta = c_\beta dH\sqrt{\iota}$ in Algorithm 1 with $\iota = 2\log(4dHN/\delta)$, then with probability at least $1 - \delta$, StepMix-LSVI (Algorithm 1) simultaneously (i) satisfies the conservative*

constraint in Equation (2)*, and (ii) achieves a total regret that is at most*

$$c_1\sqrt{d^3H^4N\iota^2} + \frac{c_2 d^3 H^4 \Delta_0 \iota^2}{\kappa^2}, \tag{6}$$

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

Remark 1. Theorem 1 indicates that StepMix-LSVI achieves a regret in the order of $\tilde{O}(\sqrt{d^3H^4N})$, the same as LSVI-UCB in Jin et al. (2020), while ensuring zero constraint violation with high probability. The conservative exploration only leads to an additive constant term in the learning regret bound in Equation (6). Besides, when $\gamma = 0$, the algorithm reduces to standard LSVI-UCB, and the additive constant term becomes zero.

The proof of Theorem 1 is provided in Appendix E where important lemmas can be found in Appendix D. We outline the major steps of the proof as follows.

First, we note that the mixture policy π^n is stochastic in general, as opposed to the deterministic greedy policy under LSVI-UCB. To cope with the policy randomness and temporal dependency, we develop a new uniform concentration lemma for value functions under policy $\tilde{\pi}^{n,h_0}$ for any $h_0 \in [0 : H+1]$, as elaborated in Lemma 7. Thus, the uniform concentration can be established for any mixture of $\tilde{\pi}^{n,h_0}$ and $\tilde{\pi}^{n,h_0-1}$. Such uniform concentration ensures that with high probability, the true value functions are bounded by the constructed UCB and LCB in Algorithm 1 (see Lemma 9). Thus, when the LCB of a policy is above the threshold γ, it ensures its safety with high probability. Moreover, the gap between them is controlled by the total expected bonus within an episode, i.e. $\sum_{h=1}^H \mathbb{E}_{\pi^n}\left[\|\phi(x_h,a_h)\|_{(\Lambda_h^n)^{-1}}\right]$, where π^n is the actual policy executed in episode n, chosen from π^b, $\bar{\pi}^n$, and $\tilde{\pi}^{n,h_n}$ (see Lemma 10). The next step is thus to bound the total expected bonus under π^n. However, the various forms policies π^n may choose from make our analysis significantly harder than the original analysis in Jin et al. (2020). We highlight several major challenges in the following.

First, when π^n is a step mixture policy, we do not have a direct estimation on the corresponding value function. Thanks to the *one-step-difference* of $\tilde{\pi}^{n,h_n}$ and $\tilde{\pi}^{n,h_n-1}$, the true value function under the step mixture policy π^n is a *linear combination* of $V^{\tilde{\pi}^{n,h_n}}$ and $V^{\tilde{\pi}^{n,h_n-1}}$ (see Lemma 1). The linearity also holds for the LCB, which ensures the **safety** of the π^n for each episode n. Besides, the difference between the LCB and V^{π^n} is controlled by $\sum_{h=1}^H \mathbb{E}_{\pi^n}\left[\|\phi(x_h,a_h)\|_{(\Lambda_h^n)^{-1}}\right]$, similar to that between the UCB and true value function under the optimistic policy $\bar{\pi}^n$ (see Lemma 11).

Another challenge is due to the randomness of the step mixture policy. Since the actions taken under the same step mixture policy may be different, the information matrix Λ_h^n may have different realizations. To cope with such randomness, we relate the information matrix with its expectation, i.e., $\bar{\Lambda}_h^n := \lambda I + \sum_{\tau=1}^{n-1} \mathbb{E}_{\pi^\tau}\left[\phi(x_h,a_h)\phi(x_h,a_h)^\top\right]$, and show that the elliptical potential $\|\phi(x_h,a_h)\|_{(\Lambda_h^n)^{-1}}$ is upper bounded by $\|\phi(x_h,a_h)\|_{(\bar{\Lambda}_h^n)^{-1}}$ up to a constant factor (see Lemma 5).

In order to bound the **regret**, we first relate the step mixture policy with the baseline policy. Intuitively, since $\bar{\pi}^n$ is an optimistic policy, switching from π^b to $\bar{\pi}^n$ after step h_n increases the value function in general. Thus, V^{π^n}, after padding the bonus terms, should be larger than V^{π^b}. Combining with the gap between LCB and its true value, we have the following corollary.

Corollary 1 (Informal)**.** *Denote* $\underline{V}^{\pi^n}$ *as the LCB of* V^{π^n}*. With probability at least* $1-\delta$*, we have*

$$V^{\pi^b} \le \underline{V}^{\pi^n} + 4\beta \sum_{h'=1}^H \mathbb{E}_{\pi^n}\left[\|\phi(x_{h'},a_{h'})\|_{(\Lambda_h^n)^{-1}}\right]. \tag{7}$$

We also note that the LCB of V^{π^n} when π^n is a step mixture policy is exactly equal to the threshold γ. Therefore,

$$V^{\pi^b} - \gamma \le 4\beta \sum_{h'=1}^H \mathbb{E}_{\pi^n}\left[\|\phi(x_{h'},a_{h'})\|_{(\Lambda_h^n)^{-1}}\right]. \tag{8}$$

Let $\mathcal{N}$ be the subset of episodes when a step mixture policy is adopted. Summing Equation (8) over $\mathcal{N}$, we have the right hand side bounded by $\tilde{O}(\sqrt{|\mathcal{N}|})$ due to a revised elliptical potential lemma (see Lemma 4). However, the left hand side equals $\kappa|\mathcal{N}|$, which implies $\mathcal{N}$ is a finite set. Similar argument applies to π^b. Therefore, StepMix-LSVI always plays $\bar{\pi}^n$ except for a finite number of episodes, which results in adding only a constant term to the original regret $\tilde{O}(\sqrt{d^3H^4N})$.

4 THE EPSMIX-LSVI ALGORITHM

In this section, we propose an algorithm named EpsMix-LSVI to learn the optimal policy subject to the conservative constraint. Different from StepMix-LSVI in Algorithm 1, EpsMix-LSVI 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 2 (Episodic Mixture Policy)**.** Given two policies π^1 and π^2 with a parameter $\rho \in (0, 1)$, the episodic mixture policy, denoted by $\rho\pi^1 \oplus (1-\rho)\pi^2$, randomly picks π^1 with probability ρ and π^2 with probability $1-\rho$ at the beginning of an episode and plays it for the entire episode.

4.1 ALGORITHM DESIGN

The EpsMix-LSVI algorithm is presented in Algorithm 4 in Appendix C.3, and it proceeds as follows. Similar to StepMix-LSVI, at the beginning of each episode n, it first constructs an optimistic policy based on the LSVI-UCB subroutine, denoted as $\bar{\pi}^n$. It then evaluates the LCB of the expected total rewards under both $\bar{\pi}^n$ and π^b, denoted as $\underline{V}^{\bar{\pi}^n}$ and $\underline{V}^{\pi^b}$, respectively. If both $\underline{V}^{\bar{\pi}^n}$ and $\underline{V}^{\pi^b}$ are above the threshold γ, it indicates that the optimistic policy $\bar{\pi}^n$ satisfies the conservative constraint with high probability. The learner thus executes $\bar{\pi}^n$ in the following episode n. Otherwise, if $\underline{V}^{\pi^b}$ is above the threshold while $\underline{V}^{\bar{\pi}^n}$ is not, it constructs an episodic mixture policy $\rho_n\bar{\pi}^n \oplus (1-\rho_n)\pi^b$ as in Equation (10) and Equation (11) of Algorithm 4, where ρ_n is determined based on $\underline{V}^{\bar{\pi}^n}$, $\underline{V}^{\pi^b}$ and the threshold γ. It can be shown that the episodic policy satisfies the conservative constraint with high probability while balancing the exploration-exploitation tradeoff.

4.2 THEORETICAL ANALYSIS

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

Theorem 2. *There exist absolute constants c', c_β, c_3 and c_4 such that, for any $\delta \in (0, 1)$, if we choose $\lambda = c'd\log(dNH/\delta)$ and $\beta = c_\beta dH\sqrt{\iota}$ in Algorithm 4 with $\iota = 2\log(4dHN/\delta)$, then with probability at least $1-\delta$, EpsMix-LSVI (Algorithm 4) simultaneously achieves (i) the conservative guarantee in Equation* (2)*, and (ii) a total regret that is at most*

$$c_3\sqrt{d^3H^4N\iota^2} + \frac{c_4 d^3H^4\Delta_0\iota^2}{\kappa^2}. \tag{9}$$

Remark 2. Theorem 2 indicates that EpsMix-LSVI achieves a regret in the order of $O(\sqrt{d^3H^4N})$ while ensuring that the conservative constraint is satisfied with high probability. It can be observed that Equation (9) has the same form as Equation (6) except for the constants (c_3 and c_4 replacing c_1 and c_2 respectively), suggesting that both StepMix-LSVI and EpsMix-LSVI have very similar regret performances and constraint guarantees. At the same time, we note that EpsMix-LSVI is less conservative than StepMix-LSVI in the sense that, the expected total return under a *selected* policy in an episode may be below the threshold when $\underline{V}^{\bar{\pi}_n} < \gamma$. 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 $1-\delta$.

The proof of Theorem 2 is deferred to Appendix F. A sketch of the proof is as follows: First, we establish a uniform concentration of the value functions under π^n in each episode, following similar approaches as in the proof of Theorem 1. We then show that the total number of episodes where the algorithm executes π^b or the episodic mixture policy is bounded, conditional on the uniform concentration of the value functions. This ensures that the performance degradation compared with LSVI-UCB (Jin et al., 2020) is bounded. Besides, the uniform concentration also ensures that the conservative constraint is satisfied in each episode.

5 FROM BASELINE POLICY TO OFFLINE DATASET

Both EpsMix-LSVI and StepMix-LSVI critically depend on the baseline policy π^b to achieve the desired conservative guarantee. In reality, however, a baseline policy that provably satisfies the

conservative constraint may not be explicitly given to the algorithm. Instead, the learning agent may have access to an offline dataset that is collected from the target environment, 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-LSVI or StepMix-LSVI. The challenge, however, is that instead of having full confidence in the conservative guarantee of π^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, Theorem 3 states that for StepMix-LSVI, 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 3. *Let π^{off} be the output of the PEVI algorithm (Jin et al., 2021) (see Algorithm 5 in Appendix G) with $N_1 = \tilde{\Theta}(\frac{d^3H^4}{\kappa^2})$[1] offline trajectories and parameters chosen properly. If we replace the baseline policy π^b used in Algorithm 1 by π^{off}, then there exist two constants c_1, c_2 such that with probability at least $1-2\delta$, we can simultaneously (i) satisfy the conservative constraint in Equation* (2)*, and (ii) achieve a total regret that is at most*

$$c_1\sqrt{d^3H^4N\iota^2} + \frac{4c_2d^3H^4(\Delta_0+\kappa/2)\iota^2}{\kappa^2}.$$

A similar result for EpsMix-LSVI can be established, and is given as Theorem 8 in Appendix G. We see that N_1 scales inversely proportional to κ^2, suggesting that a good baseline policy would require small amount of data and vice versa. Besides, the additive term increases compared with Theorem 1. In general, a large N_1 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 N. We also note that although both Theorem 3 and Theorem 8 depend on using PEVI 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; see Appendix G for more discussion.

6 Experimental Results

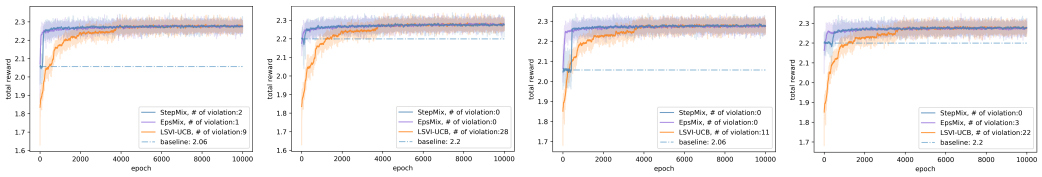
Synthetic Environment. We generate a synthetic environment to evaluate the proposed algorithms. We set the number of states $|\mathcal{S}|$ to be 10, the number of actions $|\mathcal{A}|$ for each state to be 100, and the dimension of the feature d to be 5. The feature vector $\phi(\cdot,\cdot)$ for each state-action pair is generated independently and uniformly at random from the d-dimension simplex. We also generate a $d \times |\mathcal{S}|$ matrix $\boldsymbol{\mu}$ where each row is a probability measure randomly drawn from a $|\mathcal{S}|$-dimensional simplex. Let $\{\mu(s)\}_s$ be the columns of $\boldsymbol{\mu}$. Besides, we also randomly draw θ_h uniformly from $[0,1]^d$. Such procedure guarantees that the synthetic environment is a linear MDP with rewards lying in $[0,1]$.

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

Results. We first evaluate the proposed StepMix-LSVI, EpsMix-LSVI, and compare with LSVI-UCB. We set $\lambda = 1$, $\gamma = 0.7V^{\pi^b}$. For each algorithm, we run 10 trials and plot the average results.

In Figure 1, we track the total reward obtained in each episode with different confidence bound coefficient β and baseline parameters. We also track the total number of episodes during which the total reward is below the threshold γ. We emphasize that this is different from violating the conservative constraint, as our constraint in Equation (2) is defined in terms of the *expected* total reward rather than the actual return. We have the following observations: First, StepMix-LSVI and EpsMix-LSVI converge to the optimal policy with zero or very few violations. Between them, EpsMix-LSVI appears to be more aggressive, leading to faster convergence and a few more violations. Meanwhile, LSVI-UCB also converges to the optimal policy, but with a much slower rate and much more violations. Second, the performance of baseline affects the performances of StepMix-LSVI and EpsMix-LSVI significantly. A better baseline policy leads to faster convergence. Third, the

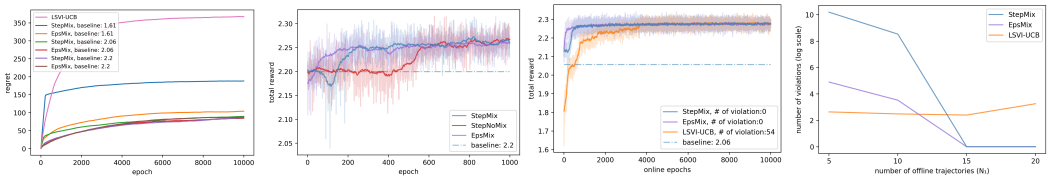
[1] We hide the logarithm factor for simplicity.



(a) $k = 20, \beta = 1$ (b) $k = 40, \beta = 1$ (c) $k = 20, \beta = 2$ (d) $k = 40, \beta = 2$

Figure 1: Total reward of each episode under StepMix-LSVI, EpsMix-LSVI, and LSVI-UCB with different β and baseline parameter k. Numbers of violations are stated in the legend.

confidence bound coefficient β also affects the performances of our algorithms, especially StepMix-LSVI, dramatically. With large β, StepMix-LSVI tends to stay with the baseline policy for more episodes, i.e., being more conservative. This is due to the fact that our conservative exploration strategy becomes too pessimistic with large β.



(a) Regret with $\beta = 1$ (b) $k = 20, \beta = 1$ (c) $k = 20, \beta = 2$, 30 offline trajectories (d) violations vs. number of offline trajectories

Figure 2: Online: (a) regret, (b) compare with StepNoMix. Offline: (c) large N_1, (d) varying N_1.

We compare the regret performances in Figure 2(a). All algorithms achieve sub-linear regret, and the regret under StepMix-LSVI and EpsMix-LSVI are much lower than that under LSVI-UCB. In general, a better baseline policy leads to lower regret under our algorithms, which is consistent with the theory. In Figure 2(b), we compare with an additional algorithm termed StepNoMix, a modified version of StepMix-LSVI. Instead of constructing a step mixture policy, StepNoMix will execute the first concatenated policy whose LCB is above the threshold. Compared with StepMix-LSVI, StepNoMix stays at the baseline for more episodes, indicating that it is less effective in exploring unknown dimensions.

Finally, we report the performance of learning from an offline dataset. When the number of offline trajectories is sufficiently large, Figure 2(c) shows that learning a baseline policy from the offline dataset and using it as an input to StepMix-LSVI and EpsMix-LSVI do not affect their performances – we observe similar reward behaviors and similar violations (i.e., numbers of episodes when the reward is below γ) as in Figure 1(c). When there are not sufficient trajectories in the offline dataset, however, we see varying degrees of violations in Figure 2(d). More results from the offline dataset experiments can be found in Appendix G.3.

7 Conclusions

We have investigated conservative exploration in episodic MDPs with linear function approximation. Different than majority of the existing literature, we considered a stringent conservative constraint that the expected total reward of the learning policy be not much worse than that of a baseline policy in every episode. This constraint has motivated us to incorporate mixture policies in conservative exploration. We proposed two LSVI-based algorithms, one with step mixture policies and the other with episodic mixture policies and randomization. Both algorithms were proved to achieve the same regret order as the constraint-free LSVI-UCB algorithm, while never violating the conservative 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.

REFERENCES

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20095–20107. Curran Associates, Inc., 2020.

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

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

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

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Nir Baram, Guy Tennenholtz, and Shie Mannor. Maximum entropy reinforcement learning with mixture policies. *arXiv preprint arXiv:2103.10176*, 2021.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14 (65):3207–3260, 2013.

Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. *Mach. Learn.*, 22(1–3):33–57, jan 1996. ISSN 0885-6125.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. *Advances in Neural Information Processing Systems*, 33, 2020.

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

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

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

Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Improved algorithms for conservative exploration in bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3962–3969, 2020a.

Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Conservative exploration in reinforcement learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 1431–1441, 26–28 Aug 2020b.

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

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Provably efficient model-free constrained rl with linear function approximation. *arXiv preprint arXiv:2206.11889*, 2022.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 4171–4180. PMLR, 2021.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 09–12 Jul 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

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

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *arXiv preprint arXiv:1611.06426*, 2016.

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

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In Kamalika Chaudhuri and Ruslan Salakhutdinov (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.

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

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34: 22931–22942, 2021.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2377–2386, New York, New York, USA, 20–22 Jun 2016. PMLR.

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 2827–2835. PMLR, 2021.

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

Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In *International Conference on Machine Learning*, pp. 307–315. PMLR, 2013.

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual optimization: Stochastically constrained Markov decision processes with adversarial losses and unknown transitions. *arXiv preprint arXiv:2003.00660*, 2020.

Nived Rajaraman, Yanjun Han, Lin Yang, Jingbo Liu, Jiantao Jiao, and Kannan Ramchandran. On the value of interaction and function approximation in imitation learning. *Advances in Neural Information Processing Systems*, 34:1325–1336, 2021.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613. PMLR, 2020.

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

Adith Swaminathan and Thorsten Joachims. 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.

Aviv Tamar, Dotan Di Castro, and Shie Mannor. 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.

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

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. 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.

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

Matteo Turchetta, Andrey Kolobov, Shital Shah, Andreas Krause, and Alekh Agarwal. Safe reinforcement learning via curriculum induction. *arXiv preprint arXiv:2006.12136*, 2020.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. *arXiv preprint arXiv:2110.04652*, 2021.

Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2021.

Honghao Wei, Xin Liu, and Lei Ying. 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.

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

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pp. 1254–1262. PMLR, 2016.

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

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10746–10756, Virtual, 13–18 Jul 2020. PMLR.

Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Accelerating safe reinforcement learning with constraint-mismatched baseline policies. In *International Conference on Machine Learning*, pp. 11795–11807. PMLR, 2021a.

Yunchang Yang, Tianhao Wu, Han Zhong, Evrard Garcelon, Matteo Pirotta, Alessandro Lazaric, Liwei Wang, and Simon S Du. A unified framework for conservative exploration. *arXiv preprint arXiv:2106.11692*, 2021b.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020.

Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. *arXiv preprint arXiv:2103.12923*, 2021.

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

Han Zhong, Zhuoran Yang, and Zhaoran Wang Csaba Szepesvári. Optimistic policy optimization is provably efficient in non-stationary mdps. *arXiv preprint arXiv:2110.08984*, 2021.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted MDPs with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 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, which is the focus of our paper, 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., 2020b; Yang et al., 2021b). Garcelon et al. (2020b) introduces a conservative upper-confidence bound for reinforcement learning (CUCRL2) algorithm for both finite horizon and average reward problems with $O(\sqrt{T})$ regret. Conservation exploration for low-rank MDPs is studied in Yang et al. (2021b) where a generic BudgetFirst algorithm instantiated with LSVI-UCB of Jin et al. (2020) is shown to match the regret upper bound of the non-conservative counterpart. We note, as discussed in Section 1, that our constraint is more stringent than these papers. Correspondingly, the algorithms and the regret analysis are also different from the prior works.

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); 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 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 (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). The constraint considered in the aforementioned papers is w.r.t. the cumulative expected cost over a horizon falling below a certain threshold, which is different than ours. In addition, constraints such as minimizing the variance (Tamar et al., 2012) or generally, maximizing some utility function of state-action pairs (Ding et al., 2021), have been investigated. A recent work (Amani et al., 2021) also studies conservative exploration with linear function approximation, and the constraint is defined using an (unknown) linear cost function of each state and action pair. Lastly, Yang et al. (2021a) studies constrained reinforcement learning with a baseline policy that may not satisfy the given set of constraints.

Linear Function Approximation. Jin et al. (2020) presents a Least-Squares Value Iteration (LSVI)-based algorithm and shows that it achieves $\tilde{\mathcal{O}}(\sqrt{d^3H^3T})$ regret, where d is the ambient dimension of feature space, H is the length of each episode, and T is the total number of steps. Importantly, such regret is independent of the number of states and actions. Yang & Wang (2020) proposes an online RL algorithm, namely the MatrixRL, that leverages ideas from linear bandit to learn a low-dimensional representation of the probability transition model while carefully balancing the exploitation-exploration tradeoff. It shows that MatrixRL achieves a regret bound $O\big(H^2 d \log T\sqrt{T}\big)$. Wang et al. (2021) proposes the USVI-UCB algorithm under a weaker optimistic closure assumption, which achieves an $\tilde{\mathcal{O}}(\sqrt{d^3H^3T})$ regret. This result is improved to $\tilde{\mathcal{O}}(d\sqrt{H^3T})$ in Zanette et al. (2020), which proposes another weaker assumption called low inherent Bellman error. Instance-dependent logarithmic regret bounds are established for linear MDPs in He et al. (2021). In addition, there is another related line of works focusing on linear mixture MDPs (Ayoub et al., 2020; Cai et al., 2020; Zhou et al., 2021). We note that those algorithms do not consider any conservative constraints into their formulation.

Safe Bandits. Bandits problem is a standard RL problem while 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 check whether the policy satisfies the conservative constraint. Kazerouni et al. (2016); Garcelon

et al. (2020a); 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) considers conservative exploration with a sample-path constraint on the actual observed rewards rather than in expectation.

B NOTATIONS

We summarize the notations frequently used in the appendix as follows. Each policy π contains H step-wise policies, i.e. $\pi := \{\pi_1, \ldots, \pi_H\}$. As standard in the literature, we use $x_h \sim (P, \pi)$ to denote a state sampled by executing policy π under the transition kernel P for $h-1$ steps. We use the notation $\mathbb{E}_{(x_h, a_h)\sim(P,\pi)}[\cdot]$ to denote the expectation over states $x_h \sim (P, \pi)$ and actions $a_h \sim \pi$. We use the short hand $\mathbb{E}_\pi[g(x_h, a_h)]$ to denote the expectation of $g(x_h, a_h)$ where x_h follows the distribution induced by the transition dynamics P and policies $\pi_1, \ldots, \pi_{h-1}$, and a_h follows the distribution π_h.

C ALGORITHMS

C.1 LSVI-UCB SUBROUTINE

We present the LSVI-UCB subroutine in the following algorithm.

Algorithm 2 LSVI-UCB Subroutine (Jin et al., 2020)

Input: $\mathcal{D}_{n-1}, \Lambda_h^n$
for step $h = H, \ldots, 1$ **do**

$$\bar{\mathbf{w}}_h^n \leftarrow (\Lambda_h^n)^{-1} \sum_{\tau=1}^{n-1} \phi^\tau \left(r_h^\tau + \max_a \bar{Q}_{h+1}^n(x_{h+1}^\tau, a) \right)$$

$$\bar{Q}_h^n(\cdot,\cdot) \leftarrow \min\left\{ (\bar{\mathbf{w}}_h^n)^\top \phi(\cdot,\cdot) + \beta \|\phi(\cdot,\cdot)\|_{(\Lambda_h^n)^{-1}}, H \right\}$$

$$\bar{\pi}_h^n(\cdot) \leftarrow \arg\max_a \bar{Q}_h^n(\cdot, a)$$

end for
Output: $\bar{\pi}^n$

C.2 LCB-V SUBROUTINE

We present the LCB-V subroutine as follows.

Algorithm 3 The LCB-V Subroutine

Input: $\pi, \mathcal{D}_n, \Lambda_h^n, \beta$
for step $h = H, \ldots, 1$ **do**

$$\underline{\mathbf{w}}_h^\pi \leftarrow (\Lambda_h^n)^{-1} \sum_{\tau=1}^{k-1} \phi^\tau \left(r_h^\tau + \mathbb{E}_{a\sim\pi_h}[\underline{Q}_{h+1}^\pi(x_{h+1}^\tau, a)] \right)$$

$$\underline{Q}_h^\pi(\cdot,\cdot) \leftarrow \max\left\{ (\underline{\mathbf{w}}_h^\pi)^\top \phi(\cdot,\cdot) - \beta \|\phi(\cdot,\cdot)\|_{(\Lambda_h^n)^{-1}}, 0 \right\}$$

end for
Output: $\underline{V}_1^\pi(x_1) = \mathbb{E}_{a\sim\pi_1}[\underline{Q}_1^\pi(x_1, a)]$

C.3 EPSMIX-LSVI ALGORITHM

We present the EpsMix-LSVI algorithm in Algorithm 4.

Algorithm 4 The EpsMix-LSVI Algorithm

Input: λ, β, π^b
$\mathcal{D}_0 \leftarrow \emptyset$
for episode $n = 1, 2, \ldots, N$ **do**
  $\Lambda_h^n \leftarrow \sum_{\tau=1}^{n-1} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^T + \lambda I$
  $\bar{\pi}^n \leftarrow$ LSVI-UCB$(\mathcal{D}_{n-1}, \Lambda_h^n)$
  $\underline{V}^{\bar{\pi}^n} \leftarrow$ LCB-V$(\bar{\pi}^n, \mathcal{D}_{n-1}, \Lambda_h^n, \beta)$
  $\underline{V}^{\pi^b} \leftarrow$ LCB-V$(\pi^b, \mathcal{D}_{n-1}, \Lambda_h^n, \beta)$
  if $\underline{V}^{\pi^b} > \gamma$ **then**
    if $\underline{V}^{\bar{\pi}^n} \geq \gamma$ **then**
      $\pi^n \leftarrow \bar{\pi}^n$
    else

$$\rho_n = \frac{\underline{V}^{\pi^b} - \gamma}{\underline{V}^{\pi^b} - \underline{V}^{\bar{\pi}^n}} \tag{10}$$

$$\pi^n \leftarrow \rho_n \bar{\pi}^n \oplus (1 - \rho_n)\pi^b \tag{11}$$

    end if
  else
    $\pi^n \leftarrow \pi^b$
  end if
  Play π^n and collect $\{(x_h^n, a_h^n, r_h(x_h^\tau, a_h^\tau))\}_{h=1}^H$
  $\mathcal{D}_n \leftarrow \mathcal{D}_{n-1} \cup \{(x_h^n, a_h^n, r_h(x_h^\tau, a_h^\tau))\}_{h=1}^H$
end for

D KEY LEMMAS

We first characterize some useful properties of the policies adopted in Algorithm 1 in the following two lemmas.

Lemma 1 (Linearity of step mixture policy)**.** *If two policies π^1 and π^2 only differ in one step h, then for any $\rho \in [0, 1]$, we have*

$$V_{h_0}^{\rho\pi^1 + (1-\rho)\pi^2}(x_{h_0}) = \rho V_{h_0}^{\pi^1}(x_{h_0}) + (1 - \rho)V_{h_0}^{\pi^2}(x_{h_0}), \forall h_0 \in [H]$$

.

Proof. It suffices to prove the case when $h_0 \leq h$ since policies π^1 and π^2 are the same after step h. Let $\pi^3 = \rho\pi^1 + (1 - \rho)\pi^2$. Suppose $\pi^i = (\pi_1^i, ..., \pi_H^i), i \in \{1, 2, 3\}$. We have $\pi_{h'}^1 = \pi_{h'}^2 = \pi_{h'}^3$ for any $h' \neq h$, and denote $\pi_{h'}^i$ by $\pi_{h'}$ when $h' \neq h$. Therefore, for any $h_0 \leq h' \leq h$, the distribution over $\mathcal{S}$ induced by $\pi_{h_0}^i, \ldots, \pi_{h'-1}^i$ and state x_{h_0} are the same across $i \in \{1, 2, 3\}$. In addition, $Q_h^{\pi^i}(x_h, a_h)$ are all the same for any $i \in \{1, 2, 3\}$. In such cases, we omit the index i, i.e. $P_{h'}^{\pi^i} := P_{h'}^\pi$ and $Q_h^{\pi^i}(x_h, a_h) = Q_h^\pi(x_h, a_h)$. Then, we can express the value function as follows

$$V_{h_0}^{\pi^i}(x) = \sum_{h'=h_0}^{h-1} \mathbb{E}_\pi[r_{h'}(x_{h'}, a_{h'})] + \mathbb{E}_{\pi_h^i}\left[Q_h^\pi(x_h, a_h)\right].$$

Since it is linear in terms of π_h, the proof is complete. □

Lemma 2. *Given an MDP and any policy π, define $\pi' = (\pi_1, \ldots, \pi_h, \pi_{h+1}^*, \ldots, \pi_H^*)$, i.e., the policy over the last $H - h$ steps are replaced by the optimal policy π^*. Then, we must have $V^\pi \leq V^{\pi'}$.*

Proof. Recall that we use $x_{h+1} \sim (P, \pi)$ to denote a state sampled by executing the policy π under the transition kernel P for h steps. Then, the difference between the two values is

$$
\begin{aligned}
&V^{\pi} - V^{\pi'} \\
&= \mathbb{E}_{x_{h+1}\sim(P,\pi), a_{h+1}\sim\pi_{h+1}}[Q^{\pi}_{h+1}(x_{h+1}, a_{h+1})] - \mathbb{E}_{x_{h+1}\sim(P,\pi), a_{h+1}\sim\pi^*_{h+1}}[Q^{*}_{h+1}(x_{h+1}, a_{h+1})] \\
&= \mathbb{E}_{x_{h+1}\sim(P,\pi)}\left[\sum_a \pi_h(a|x_{h+1}) Q^{\pi}_{h+1}(x_{h+1}, a) - \sum_a \pi^*_h(a|x_{h+1}) Q^{*}_{h+1}(x_{h+1}, a)\right] \\
&\overset{(a)}{\leq} \mathbb{E}_{x_{h+1}\sim(P,\pi)}\left[\sum_a \pi_h(a|x_{h+1}) Q^{*}_{h+1}(x_{h+1}, a) - \sum_a \pi^*_h(a|x_{h+1}) Q^{*}_{h+1}(x_{h+1}, a)\right] \\
&\overset{(b)}{=} \mathbb{E}_{x_{h+1}\sim(P,\pi)}\left[\sum_a \pi_h(a|x_{h+1}) Q^{*}_{h+1}(x_{h+1}, a) - \max_a Q^{*}_{h+1}(x_{h+1}, a)\right] \leq 0
\end{aligned}
$$

where (a) follows from the property that $Q^{\pi^*}_h(x, a) = Q^*_h(x, a) \geq Q^{\pi}_h(x, a)$ for any π, and (b) follows from the fact that the optimal policy at step h is the greedy policy w.r.t. the optimal action-value function $Q^*_h(x_h, a_h)$ (Jin et al., 2020). □

The following lemma states a standard inequality in the regret analysis for linear models in reinforcement learning and bandits problems. Refer to Lemma G.2 in Agarwal et al. (2020) and Lemma 10 in Uehara et al. (2021).

Lemma 3 (Elliptical potential lemma)**.** *Consider a sequence of $d \times d$ positive semidefinite matrices $X_1, \ldots, X_N$ with $\mathrm{tr}(X_n) \leq 1$ for all $n \in [N]$. Define $M_0 = \lambda_0 I$ and $M_n = M_{n-1} + X_n$. Then,*

$$
\sum_{n=1}^{N} \mathrm{tr}(X_n M_{n-1}^{-1}) \leq 2 \ln \det(M_N) - 2 \ln \det(M_0) \leq 2d \ln\left(1 + \frac{N}{d\lambda_0}\right).
$$

If we focus on any subset of the set $\{X_n M_{n-1}^{-1}\}_{n=1}^{N}$, we have the following revised elliptical potential lemma.

Lemma 4 (Revised elliptical potential lemma (He et al., 2021))**.** *Consider the setup of lemma 3. Then $\forall \mathcal{N} := \{n_1, \ldots, n_{|\mathcal{N}|}\} \subset [N]$,*

$$
\sum_{n\in\mathcal{N}} \mathrm{tr}(X_n M_{n-1}^{-1}) \leq 2 \ln \det\left(\sum_{n\in\mathcal{N}} X_n + M_0\right) - 2 \ln \det(M_0) \leq 2d \ln\left(1 + \frac{|\mathcal{N}|}{d\lambda_0}\right).
$$

Proof. Define $\tilde{M}_0 = \lambda_0 I$, $\tilde{M}_i = \tilde{M}_{i-1} + X_{n_i}$ for $i \geq 1$. We have

$$
\begin{aligned}
\sum_{n\in\mathcal{N}} \mathrm{tr}(X_n M_{n-1}^{-1}) &= \sum_{i=1}^{|\mathcal{N}|} \mathrm{tr}(X_{n_i} M_{n_{i-1}}^{-1}) \\
&\overset{(a)}{\leq} \sum_{i=1}^{|\mathcal{N}|} \mathrm{tr}(X_{n_i} \tilde{M}_{i-1}^{-1}) \\
&\overset{(b)}{\leq} 2 \ln \det\left(\sum_{n\in\mathcal{N}} X_n + M_0\right) - 2 \ln \det(M_0) \\
&\overset{(c)}{\leq} 2d \ln\left(1 + \frac{|\mathcal{N}|}{d\lambda_0}\right).
\end{aligned}
$$

where (a) is due to $M_{n_i - 1} \succeq \tilde{M}_{i-1}$, (b) and (c) follow from Lemma 3. □

Different from LSVI-UCB where the learner always greedily selects the action associated with the highest UCB of $Q^*_h(x, a)$ in each step, under the StepMix-LSVI algorithm, the adopted policy may not be deterministic. As a result, the information matrix Λ^n_h may have different realizations. To cope with such randomness, we adapt the concentration inequality of covariance matrix in Zanette et al. (2021) to our setting as follows.

Lemma 5 (Adapted from Lemma 39 in Zanette et al. (2021)). *Define* $\bar{\Lambda}_h^n = \lambda I + \sum_{\tau=1}^{n-1} \mathbb{E}_{\pi^\tau}\left[\phi(x_h, a_h)\phi(x_h, a_h)^\top\right]$. *If* $\lambda = C_\lambda d \log(2N/\delta)$ *for some absolute constant* C_λ, *then with probability at least* $1 - \frac{\delta}{2}$, *we have*

$$\|\phi(x,a)\|_{(\Lambda_h^n)^{-1}} \leq \sqrt{3}\,\|\phi(x,a)\|_{(\bar{\Lambda}_h^n)^{-1}}, \forall \phi(x,a), h \in [H], n \in [N].$$

Combining Lemma 5 and Lemma 4, we are able to upper bound the summation of the bonus terms, which is key to prove that Algorithm 1 only stays on π^b or any step mixture policies for a finite number of episodes.

Lemma 6 (Sublinearity of the summation of bonus terms). *For any subset* $\mathcal{N} \in [N]$, *the following inequality holds with probability at least* $1 - \delta/2$*:*

$$\sum_{n \in \mathcal{N}} \mathbb{E}_{\pi^n}\left[\|\phi(x_h^n, a_h^n)\|_{(\Lambda_h^n)^{-1}}\right] \leq \sqrt{6d|\mathcal{N}|\iota},$$

where $\iota = \log(4dNH/\delta)$.

Proof. By Lemma 5, we have

$$\sum_{n \in \mathcal{N}} \mathbb{E}_{\pi^n}\left[\|\phi(x_h^n, a_h^n)\|_{(\Lambda_h^n)^{-1}}\right] \leq \sqrt{3} \sum_{n \in \mathcal{N}} \mathbb{E}_{\pi^n}\left[\|\phi(x_h^n, a_h^n)\|_{(\bar{\Lambda}_h^n)^{-1}}\right].$$

Note that

$$\mathbb{E}_{\pi^n}\left[\|\phi(x_h^n, a_h^n)\|^2_{(\bar{\Lambda}_h^n)^{-1}}\right] = \operatorname{tr}\left(\mathbb{E}_{\pi^n}\left[\phi(x_h^n, a_h^n)\phi(x_h^n, a_h^n)^\top\right]\left(\bar{\Lambda}_h^n\right)^{-1}\right).$$

Using the definition of $\bar{\Lambda}_h^n$ and Lemma 4, we can complete the proof using Cauchy's inequality.

$$\begin{aligned}
&\sum_{n \in \mathcal{N}} \mathbb{E}_{\pi^n}\left[\|\phi(x_h^n, a_h^n)\|_{(\Lambda_h^n)^{-1}}\right] \\
&\quad \leq \sqrt{3} \sum_{n \in \mathcal{N}} \mathbb{E}_{\pi^n}\left[\|\phi(x_h^n, a_h^n)\|_{(\bar{\Lambda}_h^n)^{-1}}\right] \\
&\quad \leq \sqrt{3}\sqrt{|\mathcal{N}| \sum_{n \in \mathcal{N}} \mathbb{E}_{\pi^n}\left[\|\phi(x_h^n, a_h^n)\|^2_{(\bar{\Lambda}_h^n)^{-1}}\right]} \\
&\quad \leq \sqrt{6d|\mathcal{N}|\iota}.
\end{aligned}$$

□

Next, we present our main concentration lemma, which is critical for achieving a sublinear learning regret while ensuring that the conservative constraint is satisfied. Compared with Lemma D.4 in Jin et al. (2020), we establish the uniform concentration for value functions under policy $\bar{\pi}_n$, which is essentially the policy obtained under LSVI-UCB, as well as policy $\tilde{\pi}^{n,h_0}$.

Lemma 7 (Uniform concentration). *If* $\beta = c_\beta dH\sqrt{\iota}$ *and* $\iota = 2\log(4dNH/\delta)$, *there exists an absolute constant* C *which is independent of* c_β *such that the we can define events*

$$\mathcal{E}_1 := \left\{\forall n, h : \left\|\sum_{\tau=1}^{n-1} \phi^\tau \left[\bar{V}_{h+1}^{\bar{\pi}^n}(x_{h+1}^\tau) - \mathbb{P}_h \bar{V}_{h+1}^{\bar{\pi}^n}(x_h^\tau, a_h^\tau)\right]\right\|_{(\Lambda_h^n)^{-1}} \leq CdH\sqrt{\chi}\right\},$$

$$\mathcal{E}_2(h_0) := \left\{\forall n, h : \left\|\sum_{\tau=1}^{n-1} \phi^\tau \left[\underline{V}_{h+1}^{\tilde{\pi}^{n,h_0}}(x_{h+1}^\tau) - \mathbb{P}_h \underline{V}_{h+1}^{\tilde{\pi}^{n,h_0}}(x_h^\tau, a_h^\tau)\right]\right\|_{(\Lambda_h^n)^{-1}} \leq CdH\sqrt{\chi}\right\},$$

$\forall h_0 \in [0 : H+1]$, *where* $\chi = 2\log(4(c_\beta+1)dNH/\delta)$. *Let* $\mathcal{E} := \mathcal{E}_1 \cap \left(\cap_{h_0=0}^{H+1} \mathcal{E}_2(h_0)\right)$. *Then, under Algorithm 1,* $\mathbb{P}[\mathcal{E}] \geq 1 - \frac{\delta}{2}$.

Proof. First, by following the argument in the proof of Lemma D.4 in Jin et al. (2020), we can show that event $\mathcal{E}_1$ happens with probability at least $1 - \frac{\delta}{2(H+3)}$.

For event $\mathcal{E}_2$, it suffices to change the definition of function class in Lemma D.6 in Jin et al. (2020), and show that the ϵ-covering number of this function class has the same upper bound. Specifically, define

$$\mathcal{V} = \left\{ V(\cdot) = \max\left\{ \mathbb{E}_{a\sim\tilde{\pi}^{n,h_0}} \left[w^\top \phi(\cdot, a) - \|\phi(\cdot, a)\|_A \right], 0 \right\} \middle| \|w\| \le L, A = \beta^2 \Lambda^{-1} \right\}, \quad (12)$$

where $\beta \in [0, B]$, and the minimum eigenvalue of Λ is no less than λ.

Consider the distance function $dist(V, V') = \sup_x |V(x) - V'(x)|$. Since $\max(\cdot, 0)$ is a contraction operator and $\mathbb{E}_{a\sim\tilde{\pi}^{n,h_0}}$ is a linear operator, we have

$$\begin{aligned} dist(V, V') &\le \sup_{x,a} \left| w^\top \phi(x, a) - \|\phi(x, a)\|_A - w'^\top \phi(x, a) + \|\phi(x, a)\|_{A'} \right| \\ &\le \|w - w'\| + \sqrt{\|A - A'\|_F}, \end{aligned}$$

which is the same upper bound as in Jin et al. (2020).

We note that each candidate policy $\tilde{\pi}^{n,h_0}$ is a concatenation of the baseline policy π^b and the deterministic greedy policy in LSVI-UCB $\bar{\pi}^n$. π^b is independent of the trajectories. As for $\bar{\pi}^n$, given both w and A, it is also independent of the trajectories, as shown in Jin et al. (2020). Therefore, different from the mixture policy π^n, $\tilde{\pi}^{n,h_0}$ does not depend on the previous stochastic transition noise given both w and A, and the corresponding V defined in Equation (12) is thus independent of history as well. Therefore, we can apply the uniform concentration inequality to the ϵ-covering space similar to that in Jin et al. (2020), and conclude that the covering number of the function class $\mathcal{V}$ has the same upper bound. Thus, event $\mathcal{E}_2(h_0)$ happens with probability at least $1 - \frac{\delta}{2(H+3)}$.

Finally, applying the union bound over all events $\mathcal{E}_1$ and $\{\mathcal{E}_2(h_0)\}_{h_0}$, $\mathcal{E}$ happens with probability $1 - \frac{\delta}{2}$. □

The following two lemmas are a direct extension from Lemma B.4 in Jin et al. (2020). We provide the analysis of different parts for completeness.

Lemma 8. *When the good event $\mathcal{E}$ happens, there exists an absolute constant c_β such that the following inequalities hold for any fixed policy π:*

$$\left| \langle \phi(x, a), \underline{w}_h^{n,h_0} \rangle - Q_h^\pi(x, a) - \mathbb{P}_h(\underline{V}_{h+1}^{\tilde{\pi}^{n,h_0}} - V_{h+1}^\pi)(x, a) \right| \le \beta \|\phi(x, a)\|_{(\Lambda_h^n)^{-1}}, \quad (13)$$

$$\left| \langle \phi(x, a), \bar{w}_h^n \rangle - Q_h^\pi(x, a) - \mathbb{P}_h(\bar{V}_{h+1}^{\pi^n} - V_{h+1}^\pi)(x, a) \right| \le \beta \|\phi(x, a)\|_{(\Lambda_h^n)^{-1}}. \quad (14)$$

Proof. Recall that $\beta = c_\beta dH\sqrt{\iota}$. It suffices to choose c_β such that $4C\sqrt{\chi} \le c_\beta\sqrt{\iota}$. This gives us

$$4C\sqrt{\iota + 2\log(c_\beta + 1)} \le c_\beta\sqrt{\iota}.$$

Thus, there exists an absolute constant c_β such that the inequalities hold. □

Lemma 9 (LCB and UCB guarantees)**.** *Under Algorithm 1, we have*

$$\begin{aligned} &\underline{Q}_h^{\tilde{\pi}^{n,h_0}}(x, a) \le Q_h^{\tilde{\pi}^{n,h_0}}(x, a), \forall h_0, h, \\ &\bar{Q}_h^{\pi^n}(x, a) \ge Q_h^*(x, a), \forall h. \end{aligned}$$

Proof. We prove the results by induction. For the base case $h = H$, $\underline{Q}_H^{\tilde{\pi}^{n,h_0}}(x, a) \le Q_H^{\tilde{\pi}^{n,h_0}}(x, a)$ holds since the value function or Q-function vanishes at step $H + 1$. Suppose $\underline{Q}_{h'}^{\tilde{\pi}^{n,h_0}}(x, a) \le Q_{h'}^{\tilde{\pi}^{n,h_0}}(x, a)$ holds for $h' \ge h + 1$. By Equation (13), we have

$$\langle \phi(x, a), \underline{w}_h^{n,h_0} \rangle - Q_h^{\tilde{\pi}^{n,h_0}}(x, a) \le \mathbb{P}_h(\underline{V}_{h+1}^{\tilde{\pi}^{n,h_0}} - V_{h+1}^{\tilde{\pi}^{n,h_0}})(x, a) + \beta \|\phi(x, a)\|_{(\Lambda_h^n)^{-1}}.$$

Then, by induction and the definition of V-functions, the first term of RHS is negative, which completes the proof of the first part. The second part directly follows Lemma B.5 in Jin et al. (2020). □

Due to the stochastic nature of the step mixture policy adopted in Algorithm 1, the recursive formula in the original analysis in Jin et al. (2020) does not work here. To handle this, we keep tracking the expectations of the differences between the LCB/UCB and the corresponding true value functions instead of the empirical values.

Lemma 10 (Bound the difference between LCB/UCB and the true value functions)**.** *Let* $\underline{\delta}_h^{n,h_0} = \mathbb{E}_{x_h \sim \tilde{\pi}^{n,h_0}} \left[V_h^{\tilde{\pi}^{n,h_0}}(x_h) - \underline{V}_h^{\tilde{\pi}^{n,h_0}}(x_h) \right]$*, Similarly, for the UCB, we let* $\bar{\delta}_h^n = \mathbb{E}_{\bar{\pi}^n} \left[\bar{V}_h^{\bar{\pi}^n}(x_h) - V_h^{\bar{\pi}^n}(x_h) \right]$*. Then, conditioned on the good event* $\mathcal{E}$*, we have*

$$\underline{\delta}_h^{n,h_0} \leq \underline{\delta}_{h+1}^{n,h_0} + 2\beta \mathbb{E}_{\tilde{\pi}^{n,h_0}} \left[\|\phi(x_h, a_h)\|_{(\Lambda_h^n)^{-1}} \right],$$
$$\bar{\delta}_h^n \leq \bar{\delta}_{h+1}^n + 2\beta \mathbb{E}_{\bar{\pi}^n} \left[\|\phi(x_h, a_h)\|_{(\Lambda_h^n)^{-1}} \right].$$

Proof. First, based on Equation (13), we have

$$Q_h^\pi(x,a) - \langle \phi(x,a), \underline{w}_h^{n,h_0} \rangle + \mathbb{P}_h(\underline{V}_{h+1}^{\tilde{\pi}^{n,h_0}} - V_{h+1}^\pi)(x,a) \leq \beta \|\phi(x,a)\|_{(\Lambda_h^n)^{-1}}$$

. Then, according to the definition of $\underline{\delta}_h^{n,h_0}$, we have

$$\begin{aligned}
&\mathbb{E}_{x_h \sim \tilde{\pi}^{n,h_0}} \left[V_h^{\tilde{\pi}^{n,h_0}}(x_h) - \underline{V}_h^{\tilde{\pi}^{n,h_0}}(x_h) \right] \\
&= \mathbb{E}_{(x_h,a_h) \sim \tilde{\pi}^{n,h_0}} \left[Q^{\tilde{\pi}^{n,h_0}}(x_h,a_h) - \underline{Q}^{\tilde{\pi}^{n,h_0}}(x_h,a_h) \right] \\
&\leq \mathbb{E}_{(x_h,a_h) \sim \tilde{\pi}^{n,h_0}} \left[-\mathbb{P}_h \left(\underline{V}_{h+1}^{\tilde{\pi}^{n,h_0}} - V_{h+1}^{\tilde{\pi}^{n,h_0}} \right)(x_h,a_h) + 2\beta \|\phi(x_h,a_h)\|_{(\Lambda_h^n)^{-1}} \right] \\
&= \mathbb{E}_{(x_h,a_h) \sim \tilde{\pi}^{n,h_0}} \left[\mathbb{P}_h \left(V_{h+1}^{\tilde{\pi}^{n,h_0}} - \underline{V}_{h+1}^{\tilde{\pi}^{n,h_0}} \right)(x_h,a_h) \right] + 2\beta \mathbb{E}_{\tilde{\pi}^{n,h_0}} \left[\|\phi(x_h,a_h)\|_{(\Lambda_h^n)^{-1}} \right] \\
&= \underline{\delta}_{h+1}^{n,h_0} + 2\beta \mathbb{E}_{(x_h,a_h) \sim \tilde{\pi}^{n,h_0}} \left[\|\phi(x_h,a_h)\|_{(\Lambda_h^n)^{-1}} \right].
\end{aligned}$$

The UCB part follows a similar argument. By Equation (14) and the definition of $\bar{\delta}_h^n$, we have

$$\begin{aligned}
&\mathbb{E}_{\bar{\pi}^n} \left[\bar{V}_h^{\bar{\pi}^n}(x_h) - V_h^{\bar{\pi}^n}(x_h) \right] \\
&= \mathbb{E}_{\bar{\pi}^n} \left[\bar{Q}^{\bar{\pi}^n}(x_h^n, a_h) - Q^{\bar{\pi}^n}(x_h,a_h) \right] \\
&\leq \mathbb{E}_{\bar{\pi}^n} \left[\mathbb{P}_h \left(\bar{V}_{h+1}^{\bar{\pi}^n} - V_{h+1}^{\bar{\pi}^n} \right)(x_h,a_h) + 2\beta \|\phi(x_h,a_h)\|_{(\Lambda_h^n)^{-1}} \right] \\
&= \mathbb{E}_{\bar{\pi}^n} \left[\mathbb{P}_h \left(\bar{V}_{h+1}^{\bar{\pi}^{n,h_0}} - V_{h+1}^{\bar{\pi}^n} \right)(x_h,a_h) \right] + 2\beta \mathbb{E}_{\bar{\pi}^n} \left[\|\phi(x_h,a_h)\|_{(\Lambda_h^n)^{-1}} \right] \\
&= \bar{\delta}_{h+1}^n + 2\beta \mathbb{E}_{\bar{\pi}^n} \left[\|\phi(x_h,a_h)\|_{(\Lambda_h^n)^{-1}} \right].
\end{aligned}$$

□

Since the policy executed by StepMix-LSVI is π^n, it is necessary to characterize the performance of π^n by examining its value function. As mentioned in Section 3.2, we define the LCB of $Q_h^{\pi^n}(x,a)$, denoted by $\underline{Q}^{\pi^n}(x,a)$, as $\rho_n \underline{Q}_h^{\tilde{\pi}^{n,h_n-1}}(x,a) + (1-\rho_n)\underline{Q}_h^{\tilde{\pi}^{n,h_n}}(x,a)$. Similarly, we define LCB of $V_h^{\pi^n}(x)$ as $\underline{V}_h^{\pi^n}(x) = \rho_n \underline{V}_h^{\tilde{\pi}^{n,h_n-1}}(x) + (1-\rho_n)\underline{V}_h^{\tilde{\pi}^{n,h_n}}(x)$. We remark that this is actually the output of the LCB-V subrountine (Algorithm 3) if the input is π^n, although we do not utilize this property in our analysis. Instead, we show that the following recursive formula also holds for π^n.

Lemma 11. *Define* $\delta_h^n = \mathbb{E}_{\pi^n} \left[V_h^{\pi^n}(x_h) - \underline{V}_h^{\pi^n}(x_h) \right]$*. Then, under Algorithm 1, we have*

$$\delta_h^n \leq \delta_{h+1}^n + 2\beta \mathbb{E}_{\pi^n} \left[\|\phi(x_h,a_h)\|_{(\Lambda_h^n)^{-1}} \right].$$

Proof. By Equation (13), for any fixed policy π, we have

$$Q_h^\pi(x,a) - \underline{Q}_h^{\tilde{\pi}^{n,h_0}}(x,a) \leq \mathbb{P}_h(V_{h+1}^\pi - \underline{V}_{h+1}^{\tilde{\pi}^{n,h_0}})(x,a) + 2\beta \|\phi(x,a)\|_{(\Lambda_h^n)^{-1}}. \tag{15}$$

Then, we replace h_0 in Equation (15) by $h_n - 1$ and h_n, multiply coefficients ρ_n and $(1 - \rho_n)$ on these two inequalities, respectively, and add them together. We have

$$Q_h^{\pi}(x,a) - \underline{Q}_h^{\pi^n}(x,a) \leq \mathbb{P}_h(V_{h+1}^{\pi} - \underline{V}_{h+1}^{\pi^n})(x,a) + 2\beta \left\|\phi(x,a)\right\|_{(\Lambda_h^n)^{-1}}.$$

Finally, according the definition of δ_h^n, we have

$$\begin{aligned}
&\mathbb{E}_{\pi^n}\left[V_h^{\pi^n}(x_h) - \underline{V}_h^{\pi^n}(x_h)\right] \\
&= \mathbb{E}_{\pi^n}\left[Q^{\pi^n}(x_h,a_h) - \underline{Q}^{\pi^n}(x_h,a_h)\right] \\
&\leq \mathbb{E}_{\pi^n}\left[-\mathbb{P}_h\left(\underline{V}_{h+1}^{\pi^n} - V_{h+1}^{\pi^n}\right)(x_h,a_h) + 2\beta\left\|\phi(x_h,a_h)\right\|_{(\Lambda_h^n)^{-1}}\right] \\
&= \mathbb{E}_{\pi^n}\left[\mathbb{P}_h\left(V_{h+1}^{\tilde{\pi}^n} - \underline{V}_{h+1}^{\pi^n}\right)(x_h,a_h)\right] + 2\beta\mathbb{E}_{\pi^n}\left[\left\|\phi(x_h,a_h)\right\|_{(\Lambda_h^n)^{-1}}\right] \\
&= \delta_{h+1}^n + 2\beta\mathbb{E}_{\pi^n}\left[\left\|\phi(x_h,a_h)\right\|_{(\Lambda_h^n)^{-1}}\right].
\end{aligned}$$

□

Lemma 12. *Define $\pi^{b,h,*} = (\pi_1^b, ..., \pi_h^b, \pi_{h+1}^*, ..., \pi_H^*)$. The difference between LCB $\underline{V}^{\pi^n}$ and the true value function $V^{\pi^{b,h_n,*}}$ is bounded as follows:*

$$V^{\pi^{b,h_n,*}} \leq \underline{V}^{\pi^n} + 4\beta\sum_{h'=1}^{H}\mathbb{E}_{\pi^n}\left[\left\|\phi(x_{h'},a_{h'})\right\|_{(\Lambda_h^n)^{-1}}\right].$$

Proof. Note that for any $h_0 \in \{0,1,\ldots,H+1\}$, the difference between $\pi^{b,h_0,*}$ and $\tilde{\pi}^{n,h_0}$ is the last $H - h_0$ steps, where $\pi_{h'}^*$ is replaced by $\bar{\pi}_{h'}^n$. Therefore, we compare the two value functions $V_1^{\pi^{b,h_0,*}}$ and $V_1^{\tilde{\pi}^{n,h_0}}$ as follows:

$$\begin{aligned}
V^{\pi^{b,h_0,*}} &= V_1^{\tilde{\pi}^{n,h_0}} - \mathbb{E}_{\pi^b}\left[V_{h_0+1}^{\bar{\pi}^n}(x_{h_0+1}) - V_{h_0+1}^*(x_{h_0+1})\right] \\
&\leq V^{\tilde{\pi}^{n,h_0}} - \mathbb{E}_{\pi^b}\left[V_{h_0+1}^{\bar{\pi}^n}(x_{h_0+1}) - \bar{V}_{h_0+1}^{\bar{\pi}^n}(x_{h_0+1})\right] \\
&\leq V^{\tilde{\pi}^{n,h_0}} + \mathbb{E}_{\tilde{\pi}^{n,h_0}}\left[\sum_{h'=h_0+1}^{H} 2\beta\mathbb{E}_{\bar{\pi}^n}\left[\left\|\phi(x_{h'},a_{h'})\right\|_{(\Lambda_{h'}^n)^{-1}}\right]\right]. \qquad (16)
\end{aligned}$$

Substituting h_0 in Equation (16) with h_n and $h_n - 1$, respectively, we have,

$$V^{\pi^{b,h_n,*}} \leq V_1^{\tilde{\pi}^{n,h_n}} + \mathbb{E}_{\tilde{\pi}^{n,h_n}}\left[\sum_{h'=h_n+1}^{H} 2\beta\left[\left\|\phi(x_{h'},a_{h'})\right\|_{(\Lambda_{h'}^n)^{-1}}\right]\right] \qquad (17)$$

$$V^{\pi^{b,h_n-1,*}} \leq V_1^{\tilde{\pi}^{n,h_n-1}} + \mathbb{E}_{\tilde{\pi}^{n,h_n-1}}\left[\sum_{h'=h_n}^{H} 2\beta\left[\left\|\phi(x_{h'},a_{h'})\right\|_{(\Lambda_{h'}^n)^{-1}}\right]\right]. \qquad (18)$$

Multiplying Equation (17) and Equation (18) by ρ_n and $1-\rho_n$ respectively and adding them together, and noting that $V_1^{\pi^{b,h_n,*}} \leq V_1^{\pi^{b,h_n-1,*}}$, we have

$$\begin{aligned}
V_1^{\pi^{b,h_n,*}} &\leq V_1^{\pi^n} + 2\mathbb{E}_{\pi^n}\left[\sum_{h=h_n}^{H}\beta\left[\left\|\phi(x_h,a_h)\right\|_{(\Lambda_h^n)^{-1}}\right]\right], \\
&\leq \underline{V}_1^{\pi^n} + 4\mathbb{E}_{\pi^n}\left[\sum_{h=h_n}^{H}\beta\left[\left\|\phi(x_h,a_h)\right\|_{(\Lambda_h^n)^{-1}}\right]\right].
\end{aligned}$$

□

E PROOF OF THEOREM 1

Lemma 13. *The total expected reward received in episode n under policy π^n is always above γ under the good event $\mathcal{E}$.*

Proof. First, we note that π^n is a linear combination of two policies $\tilde{\pi}^{n,h_n}$ and $\tilde{\pi}^{n,h_n-1}$, which only differ in one step h_n. According to Lemma 1, we have

$$\begin{aligned} V^{\pi^n} &= \rho_n V^{\tilde{\pi}^{n,h_n-1}} + (1-\rho_n) V^{\tilde{\pi}^{n,h_n}} \\ &\geq \frac{\underline{V}^{\tilde{\pi}^{n,h_n}} - \gamma}{\underline{V}^{\tilde{\pi}^{n,h_n}} - \underline{V}^{\tilde{\pi}^{n,h_n-1}}} \underline{V}^{\tilde{\pi}^{n,h_n-1}} + \frac{\gamma - \underline{V}^{\tilde{\pi}^{n,h_n-1}}}{\underline{V}^{\tilde{\pi}^{n,h_n}} - \underline{V}^{\tilde{\pi}^{n,h_n-1}}} \underline{V}^{\tilde{\pi}^{n,h_n}} \\ &= \gamma. \end{aligned}$$

Note that the inequality holds with probability at least $1-\delta$, which is due to Lemma 9. □

The next two lemmas show that the algorithm would always use the policy $\bar{\pi}^n$ except for a finite number of episodes.

Lemma 14. *Conditioned on the good event $\mathcal{E}$, the number of episodes in which $h_n \in \{1, 2, \ldots, H+1\}$ is finite.*

Proof. It suffices to show the conclusion holds for each $h_n = h \in [1 : H+1]$. We first consider the case when $h_n = H+1$. Let $\mathcal{N}_{H+1} = \{n : \pi^n = \pi^b\}$. Recall that $h_n = H+1$ implies $\pi^n = \pi^b$. With the condition $\pi^n = \pi^b$, we must have $\underline{V}^{\pi_b} \leq \gamma = V^{\pi^b} - \kappa$. Therefore,

$$\begin{aligned} |\mathcal{N}_{H+1}|\kappa &\leq \sum_{n \in \mathcal{N}_{H+1}} V^{\pi_b} - \underline{V}^{\pi_b} \\ &\overset{(a)}{\leq} \sum_{n \in \mathcal{N}_{H+1}} \sum_{h=1}^{H} 2\beta \mathbb{E}_{\pi^b} \|\phi^k\|_{(\Lambda_h^n)^{-1}} \\ &\overset{(b)}{\leq} 2c_\beta dH^2 \sqrt{\iota} \sqrt{6d|\mathcal{N}_{H+1}|\iota} \\ &= 2c_\beta \sqrt{6d^3H^4|\mathcal{N}_{H+1}|\iota^2}, \end{aligned}$$

where (a) is due to Lemma 10, and (b) follows from Lemma 6. Thus, $|\mathcal{N}_{H+1}| \leq \frac{24c_\beta^2 d^3 H^4 \iota^2}{\kappa^2}$.

Then, for any $h \in \{1, 2, \ldots, H\}$, let $\mathcal{N}_h = \{n : h_n = h\}$. We have that when $h_n = h, \rho_n < 1$ according to Algorithm 1. The equality $\underline{V}^{\pi^n} = \gamma = V^{\pi^b} - \kappa$ holds. Then, we have

$$\begin{aligned} |\mathcal{N}_h|\kappa &= \sum_{n \in \mathcal{N}_h} \left(V^{\pi^b} - \underline{V}^{\pi^n} \right) \\ &\overset{(a)}{\leq} \sum_{n \in \mathcal{N}_h} \left(V^{\pi_b} - V^{\pi^{b,h,*}} + 4\beta \sum_{h' \geq h} \mathbb{E}_{\pi^n} \left[\|\phi(x_{h'}, a_{h'})\|_{(\Lambda_{h'}^n)^{-1}} \right] \right) \\ &\overset{(b)}{\leq} 4\beta \sum_{n \in \mathcal{N}_h} \sum_{h \geq 1} \mathbb{E}_{\pi^n} \left[\|\phi_h^n\|_{(\Lambda_h^n)^{-1}} \right], \end{aligned}$$

where (a) is due to Lemma 12 and (b) is due to Lemma 2.

Finally, taking the summation over all h, we have

$$\sum_{h=1}^{H+1} |\mathcal{N}_h|\kappa \leq 4\beta \sum_{n \in \cup_{h \geq 1} \mathcal{N}_h} \sum_{h \geq 1} \mathbb{E}_{\pi^n} \left[\|\phi_h^n\|_{(\Lambda_h^n)^{-1}} \right] \tag{19}$$

$$\overset{(a)}{\leq} 4c_\beta \sqrt{6d^3H^4 \left(\sum_{h \geq 1} |\mathcal{N}_h| \right) \iota^2} \tag{20}$$

where (a) follows from Lemma 6. Hence, $\sum_{h\geq 1}|\mathcal{N}_h|$ is upper bounded by $\frac{96c_\beta^2 d^3H^4\iota^2}{\kappa^2}$. □

Finally we are ready to prove Theorem 1. We restate it as follows.

Theorem 4. *There exist absolute constants c', c_β, c_1 and c_2 such that, for any $\delta \in (0,1)$, if we choose $\lambda = c'd\log(dNH/\delta)$ and $\beta = c_\beta dH\sqrt{\iota}$ in Algorithm 1 with $\iota = 2\log(4dHN/\delta)$, then with probability at least $1-\delta$, StepMix-LSVI (Algorithm 1) (i) satisfies the conservative guarantee in Equation* (2)*, and (ii) achieves a total regret that is at most*

$$c_1\sqrt{d^3H^4N\iota^2} + \frac{c_2 d^3H^5\Delta_0\iota^2}{\kappa^2},$$

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

Proof. Conditioned on the good event $\mathcal{E}$, recall that c_β is chosen based on Lemma 8. The first part is a direct result from Lemma 13. For the second part, recall that $\mathcal{N}_0 = \{n : h_n = 0\} = \{n : \pi^n = \bar{\pi}^n\}$. By Lemma 14 and Lemma 9, we have

$$\begin{aligned}
R_N &= \sum_{n\in[N]}(V^* - V^{\pi^n}) \\
&= \sum_{n\in\mathcal{N}_0}(V^* - V^{\bar{\pi}^n}) + \sum_{n\notin\mathcal{N}_0}(V^* - V^{\pi^b}) + \sum_{n\notin\mathcal{N}_0}(V^{\pi^b} - V^{\pi^n}) \\
&\overset{(a)}{\leq} \sum_{n\in\mathcal{N}_0}(\bar{V}^{\bar{\pi}^n} - V^{\bar{\pi}^n}) + \sum_{h=1}^{H+1}|\mathcal{N}_h|\Delta_0 + \sum_{n\notin\mathcal{N}_0}\sum_{h\in[H]}\mathbb{E}_{\pi^n}\left[2\beta\|\phi^n\|_{(\Lambda_h^n)^{-1}}\right] \\
&\overset{(b)}{\leq} \sum_{n\in[N]}\sum_{h\in[H]}\mathbb{E}_{\pi^n}\left[2\beta\|\phi^n\|_{(\Lambda_h^n)^{-1}}\right] + \frac{96c_\beta^2 d^3H^4\Delta_0\iota^2}{\kappa^2} \\
&\overset{(c)}{\leq} 2c_\beta\sqrt{6d^3H^4N\iota^2} + \frac{96c_\beta^2 d^3H^4\Delta_0\iota^2}{\kappa^2},
\end{aligned}$$

where (a) is due to Lemma 12 and Lemma 2, (b) follows from Lemma 10 and Lemma 14, and (c) is due to Lemma 6. By choosing $c_1 = 2\sqrt{6}c_\beta, c_2 = 96c_\beta^2$, we have

$$\mathrm{Reg}(N) \leq c_1\sqrt{d^3H^4N} + \frac{c_2 d^3H^4\Delta_0\iota^2}{\kappa^2},$$

which completes the proof. □

F Proof of Theorem 2

We follow a similar approach as the proof of Theorem 1. First, we define the good event under Algorithm 4 as $\mathcal{E}_1 \cap \mathcal{E}_2(0) \cap \mathcal{E}_2(H+1)$, where the definitions of $\mathcal{E}_1$ and $\mathcal{E}_2(h_0)$ can be found in Lemma 7.

Lemma 15. *Conditioned on the good event $\mathcal{E}$, all policies $\{\pi^n\}$ executed under Algorithm 4 are safe.*

Proof. It suffices to check whether π^n is safe when $\pi^n \neq \pi^b$. There are two cases. The first case is when $\underline{V}^n < \gamma$, and the algorithm randomly picks between $\bar{\pi}^n$ and π^b. Note that $\rho_n = \frac{V^{\pi^b} - \gamma}{V^{\pi^b} - \underline{V}^n} < 1$. By Lemma 9, we have,

$$V^{\pi^n} \geq \underline{V}^{\pi^n} = \rho_n\underline{V}^n + (1-\rho_n)\underline{V}^{\pi^b} = \gamma.$$

Thus, in expectation, the episodic mixture policy is safe.

The second case is when $\pi^n = \bar{\pi}^n$, which occurs when $\underline{V}^{\pi^n} \geq \gamma$. We again apply Lemma 9 to have

$$\underline{V}^{\pi^n} = \underline{V}^{\pi^n} \geq \gamma.$$

Therefore π^n is always safe conditioned on the good event $\mathcal{E}$. □

Lemma 16. *Let $\mathcal{N}_\rho = \{n : \pi^n$ is an episodic mixture policy$\}$, and $\mathcal{N}_b = \{n : \pi^n = \pi^b\}$. Conditioned on the good event $\mathcal{E}$, $|\mathcal{N}_b|$ and $|\mathcal{N}_\rho|$ are both finite.*

Proof. First, when $\pi^n = \pi^b$, we have

$$\begin{aligned}|\mathcal{N}_b|\kappa &\le \sum_{n\in\mathcal{N}_b} V^{\pi^b} - \underline{V}^{\pi^b} \\ &\overset{(a)}{\le} \sum_{n\in\mathcal{N}_{H+1}} \sum_{h=1}^{H} 2\beta \mathbb{E}_{\pi^b} \|\phi(x_h, a_h)\|_{(\Lambda_h^n)^{-1}} \\ &\overset{(b)}{\le} 2c_\beta \sqrt{6d^3H^4|\mathcal{N}_b|\iota^2},\end{aligned}$$

where (a) is due to Lemma 10 and (b) follows from Lemma 4. Thus, $|\mathcal{N}_b| \le \frac{24c_\beta^2 d^3 H^4 \iota^2}{\kappa^2}$.

Similarly, when $n \in \mathcal{N}_\rho$, $\rho_n \underline{V}^{\bar{\pi}^n} + (1-\rho_n)\underline{V}^{\pi^b} = \gamma$ holds for all $n \in \mathcal{N}_\rho$. We again apply Lemma 10 and Lemma 4 to have the following inequalities:

$$\begin{aligned}|\mathcal{N}_\rho|\kappa &\le \sum_{n\in\mathcal{N}_\rho} V^{\pi^b} - \gamma \\ &= \sum_{n\in\mathcal{N}_\rho} V^{\pi^b} - \rho_n \underline{V}^{\bar{\pi}^n} - (1-\rho_n)\underline{V}^{\pi^b} \\ &\le \sum_{n\in\mathcal{N}_\rho} \rho_n V^{\pi^b} - \rho_n \underline{V}^{\bar{\pi}^n} + 2\beta(1-\rho_n) \sum_{h=1}^{H} \mathbb{E}_{\pi^b}\left[\|\phi(x_h,a_h)\|_{(\Lambda_h^k)^{-1}}\right] \\ &\le 2\beta \sum_{n\in\mathcal{N}_\rho}\sum_{h=1}^{H} \left(\rho_n \mathbb{E}_{\bar{\pi}^n}\left[\|\phi(x_h,a_h)\|_{(\Lambda_h^k)^{-1}}\right] + (1-\rho_n)\mathbb{E}_{\pi^b}\left[\|\phi(x_h,a_h)\|_{(\Lambda_h^k)^{-1}}\right]\right) \\ &= 2\beta \sum_{n\in\mathcal{N}_\rho}\sum_{h=1}^{H} \mathbb{E}_{\pi^n}\left[\|\phi(x_h,a_h)\|_{(\Lambda_h^k)^{-1}}\right] \\ &\le 2c_\beta\sqrt{6d^3H^4|\mathcal{N}_\rho|\iota^2}.\end{aligned}$$

Therefore, $|\mathcal{N}_\rho| \le \frac{24c_\beta^2 d^3 H^4 \iota^2}{\kappa^2}$. □

We are ready to prove Theorem 2, which is restated as follows.

Theorem 5. *There exist absolute constants c', c_β, c_3 and c_4 such that, for any $\delta \in (0,1)$, if we choose $\lambda = c' d\log(dNH/\delta)$ and $\beta = c_\beta dH\sqrt{\iota}$ in Algorithm 1 with $\iota = 2\log(4dHN/\delta)$, then with probability at least $1-\delta$, EpsMix-LSVI (Algorithm 4) (i) satisfies the conservative guarantee in Equation* (2)*, and (ii) achieves a total regret that is at most*

$$c_3\sqrt{d^3H^4N\iota^2} + \frac{c_4 d^3H^4\Delta_0\iota^2}{\kappa^2},$$

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

Proof. Define the set $\mathcal{N}_0 = \{n : \pi^n = \pi^b\}$. Recall the definition of $\mathcal{N}_b$ and $\mathcal{N}_\rho$, and we have $[N] = \mathcal{N}_0 \cup \mathcal{N}_b \cup \mathcal{N}_\rho$. Then, conditioned on the good event $\mathcal{E}$, with probability at least $1-\delta$, we

have

$$\begin{aligned}
\mathrm{Reg}(N) &= \sum_{n=1}^{N} V^* - V^{\pi^n} \\
&= \sum_{n\in\mathcal{N}_0} \left(V^* - V^{\pi^n}\right) + \sum_{n\in\mathcal{N}_\rho} \left(V^* - V^{\pi^n}\right) + \sum_{n\in\mathcal{N}_b} \left(V^* - V^{\pi^n}\right) \\
&\overset{(a)}{\leq} \sum_{n\in\mathcal{N}_0} \left(V^* - V^{\bar{\pi}^n}\right) + \sum_{n\in\mathcal{N}_\rho} \rho_n \left(V^* - V^{\bar{\pi}^n}\right) + \sum_{n\in\mathcal{N}_b\cup\mathcal{N}_\rho} \left(V^* - V^{\pi^b}\right) \\
&= \sum_{n\in\mathcal{N}_0\cup\mathcal{N}_\rho} \left(V^* - V^{\bar{\pi}^n}\right) + \left(|\mathcal{N}_b| + |\mathcal{N}_\rho|\right)\Delta_0 \\
&\overset{(b)}{\leq} \sum_{n\in\mathcal{N}_0\cup\mathcal{N}_\rho} \mathbb{E}_{\pi^n}\left[2\beta\|\phi(x_h,a_h)\|_{(\Lambda_h^n)^{-1}}\right] + \frac{48c_\beta^2 d^3H^4\Delta_0\iota^2}{\kappa^2} \\
&\overset{(c)}{\leq} 2c_\beta\sqrt{6d^3H^4N\iota^2}) + \frac{48c_\beta^2 d^3H^4\Delta_0\iota^2}{\kappa^2},
\end{aligned}$$

where (a) is because $V^{\pi^n} = \rho_n V^{\bar{\pi}^n} + (1-\rho_n)V^{\pi^b}$ when $n \in \mathcal{N}_\rho$, (b) is due to Lemma 9 and Lemma 10, and (c) follows from Lemma 4 and Lemma 16. Finally, by choosing $c_3 = 2\sqrt{6}c_\beta$ and $c_4 = 48c_\beta^2$, we complete the proof. □

G Algorithm and Theoretical Analysis with Offline Datasets

G.1 The Pessimistic Value Iteration (PEVI) Subroutine

The PEVI subroutine in Jin et al. (2021) is given below for completeness. We set the parameters as follows. Let N_1 be the solution to the following equation:

$$N_1 = \left\lceil \frac{19200d^3H^4\log(4dHN_1/\delta)}{\kappa^2} \right\rceil.$$

Then, we set $\lambda_1 = C_\lambda d\log(2N_1/\delta)$, $\iota_1 = \log(4dHN_1/\delta)$, and $\beta_1 = 20dH\sqrt{\iota_1}$, where C_λ is specified in Lemma 5.

Algorithm 5 Pessimistic Value Iteration (PEVI) (Jin et al., 2021)

Input: $\mathcal{D}_{N_1}^{\mathrm{off}}$, λ_1, β_1
$\Lambda_h^{\mathrm{off}} \leftarrow \lambda_1\mathbf{I} + \sum_{\tau=1}^{N_1}\phi(x_h^{\tau,\mathrm{off}}, a_h^{\tau,\mathrm{off}})\phi(x_h^{\tau,\mathrm{off}}, a_h^{\tau,\mathrm{off}})^\top$
$\underline{V}_{H+1}^{\mathrm{off}} \leftarrow 0$
for step $h = H, \ldots, 1$ **do**

$$\begin{aligned}
\underline{\mathbf{w}}_h^{\mathrm{off}} &\leftarrow (\Lambda_h^{\mathrm{off}})^{-1}\sum_{\tau=1}^{N_1}\phi^\tau\left(r_h^{\tau,\mathrm{off}} + \underline{V}_{h+1}^{\mathrm{off}}(x_{h+1}^{\tau,\mathrm{off}})]\right) \\
\underline{Q}_h^{\mathrm{off}}(\cdot,\cdot) &\leftarrow \max\left\{(\underline{\mathbf{w}}_h^{\mathrm{off}})^\top\phi(\cdot,\cdot) - \beta_1\|\phi(\cdot,\cdot)\|_{(\Lambda_h^{\mathrm{off}})^{-1}}, 0\right\} \\
\pi_h^{\mathrm{off}}(\cdot) &\leftarrow \arg\max_a \underline{Q}_{h+1}^{\mathrm{off}}(\cdot,a) \\
\underline{V}_h^{\mathrm{off}}(\cdot) &\leftarrow \underline{Q}_h^{\mathrm{off}}(\cdot,\pi_h^{\mathrm{off}}(\cdot))
\end{aligned}$$

end for
Output: π^{off}

G.2 THEORETICAL ANALYSIS

We first show that π^{off} can serve as a conservative baseline policy with high probability. To this end, we introduce the following operators.

$$\mathbb{B}_h \underline{V}^{\text{off}}_{h+1}(x_h, a_h) = \mathbb{E}_{x_{h+1}\sim P}[r(x_h, a_h) + \underline{V}^{\text{off}}_{h+1}(x_{h+1})|x_h, a_h],$$
$$\hat{\mathbb{B}}_h \underline{V}^{\text{off}}_{h+1}(x_h, a_h) = \phi_h(x_h, a_h)(\Lambda^{\text{off}}_h)^{-1}\left[\sum_{\tau=1}^{N_1}\phi_h(x^\tau_h, a^\tau_h)\left(r^\tau_h + \underline{V}^{\text{off}}_{h+1}(x^\tau_{h+1})\right)\right],$$

where the matrix Λ^{off}_h and the estimated value function $\underline{V}^{\text{off}}_{h+1}$ are defined in Algorithm 5.

Following Jin et al. (2021), we present a useful lemma that establishes an upper bound of the difference of two operators defined above.

Lemma 17 (Adapted from Lemma 5.2 in Jin et al. (2021))**.** *Assume the offline dataset is collected under safe baseline policy π^b in a linear MDP $\mathcal{M}$. Let $\beta_1 = 20dH\sqrt{\iota_1}$ be the parameter for Algorithm 5, where $\iota_1 = \log(4dHN_1/\delta)$. Define*

$$\mathcal{E}^{\textit{off}} = \{|(\mathbb{B}_h \underline{V}^{\textit{off}}_{h+1})(x, a) - (\hat{\mathbb{B}}_h \underline{V}^{\textit{off}}_{h+1})(x, a)| \le \Gamma_h(x, a), \forall (x, a) \in \mathcal{S}\times\mathcal{A}, h \in [H]\},$$

where $\Gamma_h(x, a) = \beta_1 \cdot (\phi(x, a)^\top (\Lambda^{\textit{off}}_h)^{-1}\phi(x, a))^{1/2}$. Then, $\mathbb{P}[\mathcal{E}^{\textit{off}}] \ge 1 - \delta/2$.

Now we are ready to prove that π^{off} is a conservative policy with probability at least $1 - \delta$.

Lemma 18. *Let π^b be the behavior policy used to collect the offline dataset that satisfies $V^{\pi^b} = \gamma + \kappa$. If $\pi^{\textit{off}}$ is the output of Algorithm 5, then, we have $\mathbb{P}[V^{\pi^{\textit{off}}} \ge \gamma + \kappa/2] \ge 1 - \delta$.*

Proof. By Lemma 5.1 in Jin et al. (2021), based on the event $\mathcal{E}^{\text{off}}$, we have

$$0 \le (\mathbb{B}_h \underline{V}^{\text{off}}_{h+1})(x, a) - \underline{Q}^{\text{off}}_h(x, a) \le 2\beta_1 \|\phi(x, a)^\top\|_{(\Lambda^{\text{off}}_h)^{-1}}. \tag{21}$$

In addition, note that

$$V^{\pi^{\text{off}}}_h(x) - \underline{V}^{\text{off}}_h(x) = \mathbb{E}_{\pi^{\text{off}}}\left[\mathbb{B}_h(V^{\pi^{\text{off}}}_{h+1} - \underline{V}^{\text{off}}_{h+1})(x, a) + (\mathbb{B}_h \underline{V}^{\text{off}}_{h+1})(s, a) - \underline{Q}^{\text{off}}_h(x, a)\right].$$

Using the facts that $V^{\pi^{\text{off}}}_{H+1} = \underline{V}^{\text{off}}_{H+1} = 0$ and Equation (21), by induction, we conclude that $\underline{V}^{\text{off}}$ is a lower bound of $V^{\pi^{\text{off}}}$, i.e., $V^{\pi^{\text{off}}} \ge \underline{V}^{\text{off}}$. Then, based on the event $\mathcal{E}^{\text{off}}$, we can bound the difference of V^{π^b} and $V^{\pi^{\text{off}}}$ as follows.

$$\begin{aligned}
V^{\pi^b} - V^{\pi^{\text{off}}} &\overset{(a)}{\le} V^{\pi^b} - \underline{V}^{\text{off}} \\
&\overset{(b)}{=} \sum_{h=1}^{H}\mathbb{E}_{\pi^b}[\langle \underline{Q}^{\text{off}}_h(x_h, \cdot), (\pi^b - \pi^{\text{off}})(\cdot|x_h)\rangle] + \sum_{h=1}^{H}\mathbb{E}_{\pi^b}[(\mathbb{B}_h \underline{V}^{\text{off}}_h)(x_h, a_h) - \underline{Q}^{\text{off}}_h(x_h, a_h)] \\
&\overset{(c)}{\le} 2\beta_1 \mathbb{E}_{\pi^b}\left[\sum_{h=1}^{H}\|\phi(x_h, a_h)\|_{(\Lambda^{\text{off}}_h)^{-1}}\right],
\end{aligned}$$

where (a) follows from the LCB property of $\underline{V}^{\text{off}}$, (b) follows from Lemma A.1 in Jin et al. (2021), and (c) is due to that π^{off} is a greedy policy and Equation (21).

By introducing the expected covariance matrix

$$\bar{\Lambda}^{\text{off}}_h = \mathbb{E}[\Lambda^{\text{off}}_h] = \lambda_1 \mathbf{I} + N_1 \mathbb{E}_{\pi^b}[\phi(x^\tau_h, a^\tau_h)\phi(x^\tau_h, a^\tau_h)^\top],$$

we can further bound $\mathbb{E}_{\pi^b}\left[\sum_{h=1}^{H}\|\phi(x_h, a_h)\|_{(\Lambda^{\text{off}}_h)^{-1}}\right]$ as follows.

$$\begin{aligned}
\mathbb{E}_{\pi^b}\left[\sum_{h=1}^{H}\|\phi(x_h,a_h)\|_{(\Lambda_h^{\text{off}})^{-1}}\right] &\overset{(a)}{\leq} \sqrt{3}\mathbb{E}_{\pi^b}\left[\sum_{h=1}^{H}\|\phi(x_h,a_h)\|_{(\bar{\Lambda}_h^{\text{off}})^{-1}}\right] \\
&\overset{(b)}{\leq} \sqrt{3}\sum_{h=1}^{H}\sqrt{\mathbb{E}_{\pi^b}\|\phi(x_h,a_h)\|^2_{(\bar{\Lambda}_h^{\text{off}})^{-1}}} \\
&= \sqrt{3}\sum_{h=1}^{H}\sqrt{\text{Tr}\left(\mathbb{E}_{\pi^b}[\phi(x_h,a_h)\phi(x_h,a_h)^\top](\bar{\Lambda}_h^{\text{off}})^{-1}\right)} \\
&\leq \sqrt{3}H\sqrt{\frac{d}{N_1}},
\end{aligned}$$

where (a) is due to Lemma 5 and (b) follows from Jensen's inequality. Note that our choice of N_1 ensures that $N_1 \geq \frac{48\beta_1^2H^2d}{\kappa^2}$. Finally ,we have

$$V^{\pi^b} - V^{\pi^{\text{off}}} \leq 2\sqrt{3}\beta_1 H\sqrt{\frac{d}{N_1}} \leq \frac{\kappa}{2}.$$

Combining the facts that $V^{\pi^b} = \gamma + \kappa$, $\mathbb{P}[\mathcal{E}^{\text{off}}] \geq 1 - \delta/2$, and Lemma 5, we have $\mathbb{P}\left[V^{\pi^{\text{off}}} \geq \gamma + \kappa/2\right] \geq 1 - \delta$. □

The next theorem generalizes our main result to the case when the baseline policy is only guaranteed to be conservative with high probability. We note that this theorem allows for more general offline algorithms and even imitation learning, including Yin et al. (2022) and Rajaraman et al. (2021), being adopted to learn a baseline policy from the offline dataset.

Theorem 6. *Given $\delta, \delta_0 \in (0,1)$, if there exists an algorithm which outputs a baseline policy $\bar{\pi}^b$ such that with probability at least $1-\delta_0$, $V^{\pi^b} \geq \gamma + \kappa_0$, then, with an overall probability at least $1 - \delta - \delta_0$, using this $\bar{\pi}^b$ instead of π^b, StepMix or EpsMix algorithm can simultaneously (i) satisfy the conservative constraint in Equation* (2)*, and (ii) achieve regrets at most*

$$c_1\sqrt{d^3H^4N\iota^2} + \frac{c_2d^3H^4(V^* - \gamma - \kappa_0)\iota^2}{\kappa_0^2} \text{ ,or , } c_3\sqrt{d^3H^4N\iota^2} + \frac{c_4d^3H^4(V^* - \gamma - \kappa_0)\iota^2}{\kappa_0^2}$$

with the same parameters as in Theorem 1 and Theorem 2.

Proof. We prove the result for StepMix, and note that the analysis for EpsMix follows the same idea.

Let $\text{Regret}_{\text{step}}(N)$ be the regret achieved by StepMix after N episodes. Define good events $\mathcal{E}_0 = \{V^{\bar{\pi}^b} \geq \gamma + \kappa_0\}$, and

$$\mathcal{E}_{\text{step}} = \left\{\text{Regret}_{\text{step}}(N) \leq c_1\sqrt{d^3H^4N\iota^2} + \frac{c_2d^3H^4(V^* - \gamma - \kappa_0)\iota^2}{\kappa_0^2} \text{ with zero constraint violation}\right\}.$$

Then, we have $\mathbb{P}[\mathcal{E}_0] \geq 1 - \delta_0$, and $\mathbb{P}[\mathcal{E}_{\text{step}}|\mathcal{E}_0] \geq 1 - \delta$, which is due to Theorem 1. Therefore,

$$\mathbb{P}[\mathcal{E}_{\text{step}}] \geq \mathbb{P}[\mathcal{E}_{\text{step}}|\mathcal{E}_0]\mathbb{P}[\mathcal{E}_0] \geq (1-\delta)(1-\delta_0) \geq 1 - \delta - \delta_0, \tag{22}$$

which completes the proof. □

Combining Lemma 18 and Theorem 6 immediately proves Theorem 3, which is restated as follows.

Theorem 7. *If we replace the baseline policy π^b used in Algorithm 1 by π^{off}, which is the output of Algorithm 5, then with probability $1 - 2\delta$, we can simultaneously (i) satisfy the conservative constraint in Equation* (2)*, and (ii) achieve a total regret that is at most*

$$c_1\sqrt{d^3H^4N\iota^2} + \frac{4c_2d^3H^4(\Delta_0 + \kappa/2)\iota^2}{\kappa^2},$$

where c_1 and c_2 are absolute constants that are the same as in Theorem 1, and $\iota = 2\log(4dHN/\delta)$.

Finally, we present the complete theorem for EpsMix-LSVI with an offline dataset in Theorem 8. The proof is the same as that for Theorem 3, i.e., combining Lemma 18 and Theorem 6.

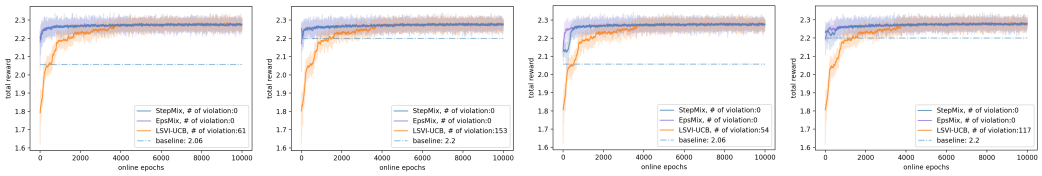
Theorem 8 (EpsMix with an offline dataset)**.** *If we replace the baseline policy π^b used in Algorithm 4 by π^{off}, which is the output of Algorithm 5, then with probability $1-2\delta$, we can simultaneously (i) satisfy the conservative constraint in Equation* (2)*, and (ii) achieve a total regret that is at most*

$$c_3\sqrt{d^3H^4N\iota^2} + \frac{4c_4d^3H^4(\Delta_0+\kappa/2)\iota^2}{\kappa^2},$$

where c_3 and c_4 are absolute constants that are the same as in Theorem 2, and $\iota = 2\log(4dHN/\delta)$.

G.3 Additional Experiments with Offline Datasets

To compare with the case of knowing the safe policy π^b in Figure 1, we adopt the same experiment settings but using 30 offline trajectories, and the results are shown in Figure 3.



(a) $k = 20, \beta = 1$ (b) $k = 40, \beta = 1$ (c) $k = 20, \beta = 2$ (d) $k = 40, \beta = 2$

Figure 3: Total reward of each episode under StepMix-LSVI, EpsMix-LSVI, and LSVI-UCB with different β and baseline parameter k with 30 offline trajectories. Numbers of violations are stated in the legend.

We see from the results that with sufficient offline trajectories, our algorithms can be safe and converge to the optimal policy. We also see that, with 30 offline trajectories, the learned policies actually have better performance than the original baseline policy. Additionally, the StepMix-LSVI and EpsMix-LSVI algorithms have a faster convergence to the optimal policy. This is because offline learning from the dataset may produce a better baseline policy (than the behavior policy), that may improve the learning performance.

H Additional Discussions

In both StepMix-LSVI and EpsMix-LSVI, a mixture policy may be adopted in some of the episodes. We note that such a *random* mixture mechanism is critical for the zero constraint violation guarantee.

To see this, we consider a special MDP where the state transition for any give state-action pair is deterministic. Besides, we also assume that the given baseline is a deterministic policy such that at each step h, π_h^b maps the current state to an action deterministically. Then, starting from the same initial state, the trajectory under π^b will always be the same. As a result, the learner is unable to get enough information outside the direction spanned by $\phi(x_h, a_h)$, where (x_h, a_h) is the fixed state-action pair at the h-th step under π^b. Therefore, if the algorithm does not allow any random mixture mechanism in the design, the learner can only do one of the following: 1) continue with π^b, which will lead to a linearly growing regret if π^b is not optimal; 2) pick another policy without mixing with π^b, which can potentially violate the conservative constraint. Such deterministic policy, therefore, may fail to simultaneously achieve sublinear regret and zero constraint violation.

Besides, we also note that the LSVI-UCB subroutine we adopt to obtain a candidate optimistic policy can be replaced by any other RL algorithms that are able to identify the optimal policy with sublinear regret in linear MDPs. Our analysis can be slightly modified to show that the corresponding regrets under StepMix-LSVI and EpsMix-LSVI remain the same order as that under the adopted RL algorithm, with additional constant terms.