A Near-Optimal Algorithm for Safe Reinforcement Learning Under Instantaneous Hard Constraints

Ming Shi [1] **Yingbin Liang** [1] **Ness Shroff** [1 2]

Abstract

In many applications of Reinforcement Learning (RL), it is critically important that the algorithm performs safely, such that instantaneous hard constraints are satisfied at each step, and unsafe states and actions are avoided. However, existing algorithms for "safe" RL are often designed under constraints that either require expected cumulative costs to be bounded or assume all states are safe. Thus, such algorithms could violate instantaneous hard constraints and traverse unsafe states (and actions) in practice. Hence, in this paper, we develop the first near-optimal safe RL algorithm for episodic Markov Decision Processes with unsafe states and actions under instantaneous hard constraints and the linear mixture model. It achieves a regret $\tilde{O}(\frac{dH^3\sqrt{dK}}{\Delta_c})$ that nearly matches the state-of-the-art regret in the setting with only unsafe actions and that in the unconstrained setting, and is safe at each step, where d is the feature-mapping dimension, K is the number of episodes, H is the episode length, and Δ_c is a safety-related parameter. We also provide a lower bound $\tilde{\Omega}(\max\{dH\sqrt{K}, \frac{H}{\Delta_c^2}\})$, which indicates that the dependency on Δ_c is necessary. Further, both our algorithm design and regret analysis involve several novel ideas, which may be of independent interest.

1. Introduction

Reinforcement learning (RL) has been extensively studied to improve the learning performance in sequential decision-making problems for machine learning applications. These decision making problems are usually modelled as a Markov Decision Process (MDP), where an online learner interacts with an unknown environment sequentially to achieve a large expected cumulative reward. Many RL algorithms that do not consider any constraint (and hence are allowed to freely explore any state-action pair) with sample-complexity guarantees have been proposed in the literature (Azar et al., 2017; Jin et al., 2018; Agarwal et al., 2019; Jin et al., 2020; Jia et al., 2020; Zhou et al., 2021b; He et al., 2022). Moreover, existing "safe" RL algorithms are usually designed under the constraint that requires expected cumulative, i.e., not *instantaneous*, costs over all steps to be bounded (Yang et al., 2019; Brantley et al., 2020; Ding et al., 2021; Paternain et al., 2022) (please see more related work in Section 1.2). Thus, practical scenarios where unsafe states and actions must be avoided at *each* time/step are not captured.

Instantaneous hard constraints are important in many practical scenarios, and any unsafe states and actions (and transitions) should be avoided at each step. In safety-critical systems, violating such a constraint could result in catastrophic consequences. For example, in power systems, it is well-known that the states of blackouts (e.g., due to violating the power-grid operation constraints) must be avoided (Amani et al., 2019; Shi et al., 2022b). In autonomous driving, improper operations that could cause dangerous states, e.g., crashing, must be avoided (Amani et al., 2021; Vamvoudakis et al., 2021). In robotics, even a single bad action could damage the machines and any undesirable state of failure must be avoided (Turchetta et al., 2016; Wachi et al., 2018).

Recently, instantaneous hard constraints have been studied in theoretical machine learning. Specifically, Amani et al. (2019) and Pacchiano et al. (2021) studied bandits with linear instantaneous constraints that require a linear safety value of the chosen action to be bounded at each step. However, it is well-known that bandits are only a very special case of MDP. Amani et al. (2021) studied safe linear MDP with linear instantaneous hard constraints. However, they still assume that only the actions could be unsafe, and hence unsafe states (and transitions) are still not considered. Intuitively, when there are only unsafe actions, any action will always lead to a state in any future step that is safe. Then, we could consider the safety at each step separately. Indeed, the existing idea in such a setting is to estimate the safe

[1] AI-EDGE Institute and Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA [2] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA. Correspondence to: Ming Shi <shi.1796@osu.edu>.

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

actions at each step separately, *without the need to consider the impact from other steps*. In sharp contrast, when one allows for the more practical scenario when unsafe states can also exist (as done in this paper), even though an action is safe at a step, it may cause unsafe states in subsequent steps. As a result, at each step, the impact from other steps must be carefully handled. This results in significantly new challenges in both the algorithm design and regret analysis.

Therefore, this paper studies an important open question: *in MDPs with unsafe states and actions (and transitions) under instantaneous hard constraints, is it possible to design an RL algorithm that not only still achieves a strong sample-complexity guarantee, but is also safe (i.e., satisfies the instantaneous hard constraint) at each step?*

1.1. Our Contributions

In this paper, we make the first effort to address this question. Specifically, we study episodic MDPs with unsafe states and actions under instantaneous hard constraints and the linear mixture model. We develop an RL algorithm, called Least-Square Value Iteration by lookiNg ahEad and peeking backWard (LSVI-NEW). LSVI-NEW achieves a regret $\tilde{O}(\frac{dH^3\sqrt{dK}}{\Delta_c})$ that nearly matches the state-of-the-art regret in the *unsafe-action* setting and that in the *unconstrained* setting, and is safe at each step, where d is the feature-mapping dimension, K is the number of episodes, H is the number of steps in each episode, and Δ_c (which is defined in Theorem 2) is a safety-related parameter. We also provide a lower bound $\tilde{\Omega}(\max\{dH\sqrt{K}, \frac{H}{\Delta_c^2}\})$, which indicates that the dependency on Δ_c is necessary.

As discussed before, in our case, the coupling between steps need to be carefully handled. To resolve the new challenges due to this coupling, our algorithm in Section 3 involves four important novel ideas. *Idea I: constructing safe subgraphs (defined in Section 2.2).* Remember that an action that is safe at a step could cause unsafe future states. To resolve this problem, we restrict LSVI-NEW to be inside safe subgraphs of the state-transition diagram. These safe subgraphs are constructed by estimating safe state-sets at each step in a backward manner, such that the chosen action could only result in future states that are estimated to be safe. *Idea II: encouraging to explore the transitions with higher uncertainty.* Due to our first idea for safety, the choices of actions become restricted. In order to still achieve a sublinear regret, the algorithm needs to be more optimistic in the learning process. To resolve this new pessimism-optimism dilemma, we construct a new bonus term in the estimated Q-value function to encourage LSVI-NEW to explore transitions with higher uncertainty. *Idea III: encouraging to explore the future subsubgraphs with higher uncertainty.* Idea-II by itself is not sufficient, since each step could be affected by the safety-learning process at future steps. For example, even though the safety function at step h may be precisely known, a bad learning quality at a future step $h' > h$ could make the algorithm still not be able to really execute the optimal safe action at step h. To resolve this difficulty, we construct another new bonus term to encourage LSVI-NEW to explore future subsubgraphs with higher uncertainty. *Idea IV: encouraging to explore the past subsubgraphs with higher uncertainty.* Similar to that in Idea III, since each step h is also affected by past steps $h' < h$, we construct a new bonus term to encourage LSVI-NEW to explore past subsubgraphs with higher uncertainty.

To show a sublinear regret of LSVI-NEW, our regret analysis involves novel ideas for solving the following difficulties. (Please see Section 4 for details.) *Difficulty I: the invariant in RL with the ergodicity property does not hold any more.* Due to our special design of safe subgraphs, the optimal policy and LSVI-NEW may visit different sets of states at each step. Thus, the classical invariant that shows the estimated V-value is larger than the optimal V-value at any state does not hold. To resolve this problem, we construct the value functions in a special way so that other useful interesting invariants still hold. *Difficulty II: how to quantify the impact from other steps?* Our idea is to consider the future and past impacts separately. Then, we could quantify such impacts based on our construction of the safe subgraphs.

1.2. Related Work

We provide more related work in this section. *To the best of our knowledge, none of existing work has addressed the fundamental open problem that we consider in this paper.*

RL with constraints: Constraints that require some expected cumulative costs over all steps to be bounded have been widely studied in safe RL (Wu et al., 2016; Achiam et al., 2017; Tessler et al., 2018; Yang et al., 2019; Efroni et al., 2020; Ding et al., 2020; Brantley et al., 2020; Kalagarla et al., 2021; Liu et al., 2021; Ding et al., 2021; Wei et al., 2021; Xu et al., 2021; Shi et al., 2022a; Paternain et al., 2022; Bai et al., 2022; Ghosh et al., 2022).

Instantaneous hard constraints with only unsafe actions: First, Amani et al. (2019); Pacchiano et al. (2021) studied safe linear bandits which require a linear safety value of the chosen action to be bounded at each step. Second, Amani et al. (2021) studied linear MDPs with instantaneous hard constraints, while assuming only actions could be unsafe. Third, another line of work focused on online optimization with instantaneous hard constraints and unsafe actions, e.g., Badiei et al. (2015); Li et al. (2020); Shi et al. (2021a;b).

Instantaneous hard constraints under deterministic transitions: Turchetta et al. (2016) and Wachi et al. (2018) studied instantaneous hard constraints with unsafe states, while assuming the state transitions are deterministic.

2. Problem Formulation

In this section, we provide the problem formulation.

2.1. Episodic MDP Under Instantaneous Hard Constraints and the Linear Mixture Model

We study the constrained episodic MDP, denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, c)$, in an online setting with K episodes, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, respectively; H denotes the number of steps in each episode; $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^{H}$, $r = \{r_h\}_{h=1}^{H}$ and $c = \{c_h\}_{h=1}^{H}$ denote the transition probability function, reward function and safety function, respectively. Let $T = HK$ denote the total number of steps. The learner interacts with the unknown environment as follows. At each step h of episode k, the learner first chooses an action $a_h^k \in \mathcal{A}$ for current state s_h^k. Then, the learner receives a reward $r_h(s_h^k, a_h^k)$, where $r_h(\cdot) : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is known. Finally, according to the *unknown* transition probability function $\mathbb{P}_h(\cdot|s_h^k, a_h^k) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, the environment draws a next state s_{h+1}^k and reveals it to the learner. Meanwhile, the learner observes a noisy safety value $\hat{c}_h^k = c_h(s_h^k, a_h^k, s_{h+1}^k) + \zeta_h^k$, where $c_h(\cdot) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is *unknown* and ζ_h^k is an additive 0-mean σ-subGaussian random variable.

Instantaneous hard constraint: At each step $h < H$ of each episode k, the following constraint must be satisfied,

$$c_h(s_h^k, a_h^k, s_{h+1}^k) \le \bar{c}, \tag{1}$$

where $\bar{c}$ is a known constant, and $c_h(s_H^k) \le \bar{c}$ must be satisfied at step H. The transition from s_h^k through a_h^k to s_{h+1}^k is said to be unsafe if constraint (1) is violated. Due to this constraint, some states and actions could also be unsafe.

- A state is said to be unsafe at step h, if there exists no action, such that constraint (1) can be satisfied, i.e., $\min_{a \in \mathcal{A}} \max_{\{s' : \mathbb{P}_h(s'|s,a)>0\}} c_h(s, a, s') > \bar{c}$.
- An action is said to be unsafe for state s at step h, if there is a non-zero probability to transit to a state, such that constraint (1) will be violated, i.e., $\max_{\{s' : \mathbb{P}_h(s'|s,a)>0\}} c_h(s, a, s') > \bar{c}$.

As discussed in Section 1, due to unsafe states and actions caused by the instantaneous hard constraint, e.g., bad movements and failures in robotics, crushing in autonomous driving and blackouts in power systems, new fundamental difficulties need to be resolved, which is the focus of this paper.

Linear mixture MDP: Due to the ergodicity under the linear function approximation $\mathbb{P}_h(\cdot|s, a) = \langle \boldsymbol{\mu}_h^*(\cdot), \boldsymbol{\phi}(s, a) \rangle$ from Jin et al. (2020), any state could be finally visited from any other state. Thus, in such a linear MDP, no algorithm can avoid the unsafe states under constraint (1). Thus, instead we borrow the linear mixture MDP model from Jia et al. (2020); Zhou et al. (2021a;b); Zhou & Gu (2022); He et al. (2022). The importance and many applications of linear mixture MDPs have been provided in these references. Specifically, the transition probability $\mathbb{P}_h(s'|s, a) = \langle \boldsymbol{\mu}_h^*, \boldsymbol{\phi}(s, a, s') \rangle$ and safety value $c_h(s, a, s') = \langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s, a, s') \rangle$ are linear functions of a given feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, where $\boldsymbol{\mu}_h^* \in \mathbb{R}^d$ and $\boldsymbol{\gamma}_h^* \in \mathbb{R}^d$ are *unknown* parameters. As typically assumed, for any bounded function $V_h : \mathcal{S} \to [0, H]$ and state-action pair (s, a), we have $\|\boldsymbol{\phi}_{V_h}(s, a)\|_2 \le D$, where $\boldsymbol{\phi}_{V_h}(s, a) = \sum_{\{s' : \mathbb{P}_h(s'|s,a)>0\}} \boldsymbol{\phi}(s, a, s') V_h(s') \in \mathbb{R}^d$. Moreover, $\|\boldsymbol{\mu}_h^*\|_2 \le L$ and $\|\boldsymbol{\gamma}_h^*\|_2 \le L$.

2.2. State-Action Subgraphs and Performance Metric

Notice that the ergodicity property, (i.e., any state could finally be visited from any other state) in classical MDPs does not hold any more under instantaneous hard constraint (1). This is because if unsafe states can be visited from any other state, it is impossible to satisfy (1) at all steps. Due to this non-ergodicity, we define two important notions below.

First, we let $\mathcal{S}_h(s, a)$ denote the set of next-states that could be transited to with non-zero probability from a state-action pair (s, a) at step h, i.e., $\mathcal{S}_h(s, a) \triangleq \{s' : \mathbb{P}_h(s'|s, a) > 0\}$.

Assumption 1. The next-state sets $\mathcal{S}_h(s, a)$ are known in advance for all h, s, a.

Note that the transition kernel $\mathbb{P}$ is still *unknown*. More importantly, Assumption 1 is necessary (even when the state transition is deterministic (Wachi et al., 2018)), since if $\mathcal{S}_h(s, a)$ is not known in advance, no safe algorithm can achieve a sub-linear regret. Specifically, (i) if an unsafe state s' that *will not be* transited to is considered for a state-action pair (s, a), the algorithm will lose the chance to explore (s, a). For example, if $\mathbb{P}(s'|s, a) = 0$ is unknown, the algorithm would never choose (s, a) to avoid the unsafe state s'. This could result in a linear-in-T regret when (s, a) is actually optimal. (ii) If an unsafe state s' that *will be* transited to is missed for (s, a), the algorithm will suffer from this unsafe state s' when choosing a at state s.

State-action subgraph: While ergodicity does not hold, an important property here is that, by executing a deterministic policy $\pi(s, h) : \mathcal{S} \times [1, H] \to \mathcal{A}$, the learner follows a closed directed state-action subgraph

$$G^{\pi} \triangleq \left\{ (s_1, \pi(s_1, 1)), \{(s_2, \pi(s_2, 2))\}_{s_2 \in \mathcal{S}_2^{\pi}}, ..., \mathcal{S}_H^{\pi} \right\},$$

where $\mathcal{S}_h^{\pi}$ denotes the set of states that are visited with non-zero probability by policy π at step h. Note that each episode ends at step H, and thus there is no further action a taken at state $s \in \mathcal{S}_H^{\pi}$. G^{π} may contain only a subset of states in state space $\mathcal{S}$. For simplicity, we assume all

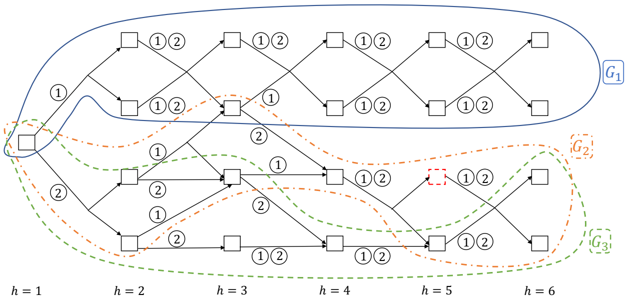

Figure 1. A sketch of subgraph examples. Squares represent states. The red dashed square at step $h = 5$ is the unsafe state. Circles represent actions. Arrows represent state transitions. There are two actions $a = 1, 2$, as shown by the numbers in the circles.

episodes start from a fixed state s_1, and drop π of G^π when it is clear from the context. In addition, we focus on the setting where the feature space of all subgraphs is convex.

Please see Figure 1 for a simple sketch of subgraph examples. For example, when choosing action $a = 1$ at all steps, the learner follows subgraph G_1. Notice that G_1 is a safe subgraph, since the unsafe state at step $h = 5$ will not be visited. As another example, the learner follows subgraph G_2 when choosing $a = 2$ at step $h = 1$, choosing $a = 1$ at step $h = 2$, choosing $a = 2$ for the second state (i.e., the second square from the top when $h = 3$) and $a = 1$ for the third state (i.e., the third square from the top when $h = 3$) at step $h = 3$, and choosing $a = 2$ at step $h = 4$ and step $h = 5$. Notice that G_2 is an unsafe subgraph, since the unsafe state at $h = 5$ could be visited. For ease of understanding, in Figure 1, we only draw finite states, two actions and three subgraphs. However, this paper considers the general linear mixture MDP with d-dimension feature mapping and convex subgraph feature space, where the number of states s, actions a and subgraphs G could be infinite.

Performance metric: We use G^{safe} to denote a safe subgraph, i.e., all state-action-state triplets (s_h, a_h, s_{h+1}) in G^{safe} satisfy the instantaneous hard constraint (1). We let $\mathcal{G}^{\text{safe}} \triangleq \{G^{\text{safe}}\}$ denote the set of all safe subgraphs. Then, the set of all possible safe *deterministic* policies is

$$\Pi^{\text{safe}} \triangleq \left\{\pi : G^\pi \in \mathcal{G}^{\text{safe}}\right\}. \tag{2}$$

Moreover, the Q-value (state-action-value) function and the V-value (state-value) function are defined as follows:

$$Q_h^\pi(s,a) \triangleq r_h(s,a) + \mathbb{E}\left[\sum_{h'=h+1}^{H} r_{h'}(s_{h'}, \pi(s_{h'}, h')) \Big| s_h = s, a_h = a\right], \tag{3}$$

$$V_h^\pi(s) \triangleq \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, \pi(s_{h'}, h')) \Big| s_h = s\right]. \tag{4}$$

Therefore, our goal is to develop an RL algorithm $\pi \triangleq \{\pi^k\}_{k=1}^K$ that (i) is safe: $\pi^k \in \Pi^{\text{safe}}$ for all k, i.e., constraint (1) is satisfied in all episodes k; (ii) achieves a sublinear regret, which is defined as

$$R^\pi \triangleq \sum_{k=1}^{K} \left\{V_1^*(s_1) - V_1^{\pi^k}(s_1)\right\}, \tag{5}$$

where $V_1^*(s_1)$ is the V-value of the optimal *safe* policy, i.e.,

$$V_1^*(s_1) = \max_{\pi \in \Pi^{\text{safe}}} V_1^\pi(s_1). \tag{6}$$

3. A Near-Optimal Safe Algorithm

In this section, we present our algorithm, called Least-Square Value Iteration by lookiNg ahEad and peeking backWard (LSVI-NEW), as shown in Algorithm 1. Before introducing our algorithm, we present a necessary assumption.

Assumption 2. (Known seed safe subgraph) There exists a known seed safe subgraph $G^{\text{safe},0} \in \mathcal{G}^{\text{safe}}$ with the known safety value c_h^0 for a state-action-state triplet $(s_h^0, a_h^0, s_{h+1}^0)$ at each step h of $G^{\text{safe},0}$.

A known seed safe subgraph is necessary for the existence of *safe* RL algorithms under instantaneous hard constraints. Without it, the unsafe states and actions cannot be avoided in the first episode. Same assumptions on such a known safe set are also made in related work (Pacchiano et al., 2021; Amani et al., 2021). As pointed out there, such an assumption is realistic since the known safe set can be obtained from existing strategies or trials with possibly low rewards.

Next, we define some notations. First, we let $\mathcal{U}_h \triangleq \{\alpha\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0) : \alpha \in \mathbb{R}\}$ denote the span of the feature $\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)$. Let $\boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}_1) \triangleq \langle \boldsymbol{\phi}_1, \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\rangle \cdot \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)$ denote the projection of a vector $\boldsymbol{\phi}_1$ to $\mathcal{U}_h$, where $\tilde{\boldsymbol{\phi}}(s, a, s') \triangleq \frac{\boldsymbol{\phi}(s,a,s')}{\|\boldsymbol{\phi}(s,a,s')\|_2}$ is the normalized vector of $\boldsymbol{\phi}(s, a, s')$. Second, we let $\mathcal{U}_h^\perp \triangleq \{\boldsymbol{\phi}_3 \in \mathbb{R}^d : \langle \boldsymbol{\phi}_3, \boldsymbol{\phi}_2\rangle = 0, \forall \boldsymbol{\phi}_2 \in \mathcal{U}_h\}$ denote the orthogonal complement of $\mathcal{U}_h$. Let $\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}_1) \triangleq \boldsymbol{\phi}_1 - \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}_1)$ denote the projection of $\boldsymbol{\phi}_1$ to $\mathcal{U}_h^\perp$. Third, we let $\boldsymbol{\phi}_{h,h+1}^k = \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k)$ denote the feature vector of the state-action-state triplet $(s_h^k, a_h^k, s_{h+1}^k)$. Let $\|\boldsymbol{x}\|_{\boldsymbol{\Lambda}} = \sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{x}}$ denote the weighted 2-norm of $\boldsymbol{x}$ with respect to $\boldsymbol{\Lambda}$. Let $\boldsymbol{I}$ denote the identity matrix.

Our LSVI-NEW algorithm contains a simple initialization phase and a more important learning phase that involves our four ideas. In the initialization phase, LSVI-NEW purely explores inside the known seed safe subgraph $G^{\text{safe},0}$, i.e., the first for-loop in Algorithm 1, where K' is a tunable parameter. This initialization phase borrows the idea in bandits with instantaneous hard constraints for obtaining and preparing some parameter information for the later learning phase (Amani et al., 2019).

Algorithm 1 Least-Square Value Iteration by lookiNg ahEad and peeking backWard (LSVI-NEW)

```
for $k = 1$ to $K'$ do
    At each step $h$, first choose the action $a_h^k = a_h(s_h^k)$ in the known seed safe subgraph $G^{\text{safe},0}$, then observe the next state $s_{h+1}^k$, finally observe the safety value $c_h(s_h^k, a_h^k, s_{h+1}^k)$.
end for
for $k = K' + 1$ to $K$ do
    for $h = H$ to $1$ do
        Step-1: Update the estimated safety parameter $\boldsymbol{\gamma}_h^k$ according to (7) and the estimated safety function $\tilde{c}_h^k$ according to (8).
        Step-2: Update the estimated safe state-set:
            $\mathcal{S}_h^{k,\text{safe}} = \{s \in \mathcal{S} | \exists a \in \mathcal{A}, \text{ s.t. (9) and (10) hold}\}$,
        and estimated safe action-set for states $s \in \mathcal{S}_h^{k,\text{safe}}$:
            $\mathcal{A}_h^{k,\text{safe}}(s) = \{a \in \mathcal{A} | \text{(9) and (10) hold for state } s\}$.
        Step-3: Update the parameter $\boldsymbol{w}_h^k$ according to (12).
        Step-4: Update the estimated $Q$-values for all state-action pairs $(s, a)$ that are estimated to be safe, i.e., $s \in \mathcal{S}_h^{k,\text{safe}}$ and $a \in \mathcal{A}_h^{k,\text{safe}}(s)$, according to (13).
    end for
    for $h = 1$ to $H - 1$ do
        Step-5: Observe the current state $s_h^k$, and then choose an action according to (17).
    end for
end for
```

From now on, we focus on introducing the five steps in the learning phase (i.e., the second for-loop in Algorithm 1) that involves four important ideas. From a high-level point of view, due to the instantaneous hard constraint, we need to carefully construct restricted safe state and action sets, such that by taking action a at state s, all subsequent steps $h' \geq h$ following (s, a) must be safe. Please see Idea I, which corresponds to Step-2 in Algorithm 1. On the other hand, because of such restrictions, the algorithm needs to learn more optimistically. Thus, we develop another three ideas for handling the impacts from current transitions, future safety and past safety, respectively. Please see Ideas II, III and IV, which correspond to Step-4 in Algorithm 1. Specifically, in Step-1, LSVI-NEW updates the regularized least-square estimator of the *projected* safety parameter $\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\gamma}_h^*)$ as follows:

$$\boldsymbol{\gamma}_h^k = (\boldsymbol{\Lambda}_{h,1}^k)^{-1} \textstyle\sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \hat{c}_h^{\tau}), \quad (7)$$

where the Gram matrix $\boldsymbol{\Lambda}_{h,1}^k = \lambda \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I}) + \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \boldsymbol{\psi}^{\mathrm{T}}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau})$, $\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I}) = \boldsymbol{I} - \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0) \tilde{\boldsymbol{\phi}}^{\mathrm{T}}(s_h^0, a_h^0, s_{h+1}^0)$, $\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \hat{c}_h^{\tau}) = \hat{c}_h^{\tau} - \frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}_{h,h+1}^{\tau}), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0) \rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0$ and $\lambda \geq d$ is a tunable parameter. Then, we estimate the safety function as follows:

$$\tilde{c}_h^k(s, a, s') = \frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}_1), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0) \rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0 + \langle \boldsymbol{\gamma}_h^k, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_1) \rangle + \beta \|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_1)\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}, \quad (8)$$

where $\boldsymbol{\phi}_1 = \boldsymbol{\phi}(s, a, s')$ and β is a tunable parameter given in Theorem 2. Notice that, on the right-hand-side (RHS) of (8), the first term is the projected safety value of (s, a, s') on $\mathcal{U}_h$, the second term is the projected empirical safety value of (s, a, s') on $\mathcal{U}_h^{\perp}$, and the last term is an upper-confidence-bound (UCB) bonus for the safety uncertainty. Thus, the accuracy of the safety value $\tilde{c}_h^k$ depends on how accurate $\boldsymbol{\gamma}_h^k$ in (7) is and how small the safety uncertainty is. Next, Step-2 in Algorithm 1 is based on $\tilde{c}_h^k$ and involves our first novel idea that is critical for guaranteeing safety.

Idea I: Constructing safe subgraphs by looking ahead. As we discussed in Section 1, in bandits and RL with only unsafe actions, the safety at each step can be estimated *separately*. In sharp contrast, due to the unsafe states and transitions in our setting, we must handle possible unsafe *future* steps. Consider Figure 1 as an example. Even though taking action $a = 1$ for the third state (the third square from the top) at step $h = 3$ is safe for $h = 3$, by doing so, the unsafe state (the red dashed square) at $h = 5$ will be visited no matter what action would be taken at $h = 4$. To resolve this new challenge, our idea is to construct special safe subgraphs where any action only results in safe future (not even just next) states. To achieve this, in Step-2, we estimate the safe state-set $\mathcal{S}_h^{k,\text{safe}}$ and action-set $\mathcal{A}_h^{k,\text{safe}}(s)$ in a *backward* manner based on the two conditions below:

$$\text{Condition 1: } \max_{s' \in \mathcal{S}_h(s,a)} \tilde{c}_h^k(s, a, s') \leq \bar{c}. \quad (9)$$

$$\text{Condition 2: } \mathcal{S}_h(s, a) \subseteq \mathcal{S}_{h+1}^{k,\text{safe}}. \quad (10)$$

Notice that, (i) condition 1 requires that by choosing action a for state s, the instantaneous hard constraint is always satisfied at step h; (ii) condition 2 requires that all possible next states in $\mathcal{S}_h(s, a)$ must be safe for next step $h + 1$. Thus, with conditions 1 and 2 satisfied simultaneously in a backward manner, all (not just next) steps $h' \geq h$ following (s, a) must be safe. Please see Theorem 1 for the safety performance of LSVI-NEW at all steps in any episode.

Moreover, since the linear mixture MDP induces a linear form of the Q-value function as follows:

$$Q_h^*(s, a) = \min\{r_h(s, a) + \langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^*}(s, a) \rangle, H\}, \quad (11)$$

in Step-3 of Algorithm 1, we update the regularized least-square estimator of the parameter w_h^* in (11) as follows:

$$\boldsymbol{w}_h^k = (\boldsymbol{\Lambda}_{h,2}^k)^{-1} \textstyle\sum_{\tau=1}^{k-1} \boldsymbol{\phi}_{h,V_{h+1}^{\tau}}^{\tau} V_{h+1}^{\tau}(s_{h+1}^{\tau}), \quad (12)$$

where the Gram matrix $\boldsymbol{\Lambda}_{h,2}^k = \lambda \boldsymbol{I} + \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_{h,V}^{\tau} \boldsymbol{\phi}_{h,V}^{\tau,\mathrm{T}}$ and $\boldsymbol{\phi}_{h,V}^{\tau} = \boldsymbol{\phi}_V(s_h^\tau, a_h^\tau)$. Then, in Step-4 of Algorithm 1, we update the Q-values of the safe state-action pairs as follows:

$$
\begin{aligned}
Q_h^k(s,a) = \min \Big\{ & H, r_h(s,a) + \langle \boldsymbol{w}_h^k, \boldsymbol{\phi}_{V_{h+1}^k}(s,a) \rangle \\
& + \epsilon_1 \cdot \|\boldsymbol{\phi}_{V_{h+1}^k}(s,a)\|_{(\boldsymbol{\Lambda}_{h,2}^k)^{-1}} \\
& + \epsilon_{h,2} \cdot \max_{s' \in \mathcal{S}_h(s,a)} \|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a,s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} \\
& + \epsilon_{h,3} \max_{(s_{h'}, a_{h'}, s') \in \mathcal{G}_h(s)} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}, a_{h'}, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}} \\
& + \epsilon_4 \max_{s' \in \mathcal{S}_1(s_1, a_1^k)} \|\boldsymbol{\psi}(\mathcal{U}_1^\perp, \boldsymbol{\phi}(s_1, a_1^k, s'))\|_{(\boldsymbol{\Lambda}_{1,1}^k)^{-1}} \Big\}, \qquad (13)
\end{aligned}
$$

where $\epsilon_1 = \beta + 1$, $\epsilon_{h,2}$, $\epsilon_{h,3}$ and ϵ_4 are given soon later, and $\mathcal{G}_h(s)$ is the set of subsubgraphs starting from state s at step h. Notice that (i) the term with ϵ_1 on the RHS of (13) is the standard Hoeffding bonus term; (ii) the terms with $\epsilon_{h,2}$, $\epsilon_{h,3}$ and ϵ_4 are three new bonus terms that we construct for capturing the impacts from future and past steps. We elaborate our novel ideas in these new bonus terms below.

Idea II: Encouraging to explore the transitions with higher uncertainty (i.e., looking ahead). As we mentioned in Section 1, there is a new pessimism-optimism dilemma in our setting. Specifically, according to the optimism-in-face-of-uncertainty principle (Azar et al., 2017), algorithms need to learn optimistically to achieve a sublinear regret. However, to avoid the unsafe states and transitions in our setting, algorithms have to be relatively pessimistic. To resolve this new dilemma, we construct a bonus term to encourage LSVI-NEW to explore the transitions with *higher* uncertainty. To achieve this, this new bonus term, i.e., the term with $\epsilon_{h,2}$ in (13), is designed to be the maximum UCB bonus over all possible next-states $s' \in \mathcal{S}_h(s,a)$.

Then, another new difficulty here is how to quantify the parameter $\epsilon_{h,2}$ for such a bonus term, such that a sublinear regret can be achieved. To resolve this problem, we set

$$
\epsilon_{h,2} = \frac{\frac{4\beta H}{\tilde{\delta}} \frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c)}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c)}}{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - \frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c)}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c)} \kappa}, \qquad (14)
$$

where $\bar{c}_{h'}^0 = \max_{h \le h' \le H} c_{h'}^0$, $\Delta_{\boldsymbol{\phi}}(c) = L \cdot \max_{s,a,h} \max_{s',s'' \in \mathcal{S}_h(s,a)} \|\boldsymbol{\phi}(s,a,s') - \boldsymbol{\phi}(s,a,s'')\|_2$, and $\tilde{\delta}$ and κ are scalars given in Theorem 2. Notice that when all states are assumed to be safe, all terms related to next state s' would be 0. Then, $\epsilon_{h,2}$ would be $\frac{4\beta H}{\bar{c} - c_h^0}$, which is the same as the parameter used in the setting with only unsafe actions (Amani et al., 2021). However, one difference here is that we need to handle the *worst* transition. Thus, the denominator needs to capture the *smallest* safety balance, i.e., $\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c)$, that is left for exploration. Another difference is that even though the safety balance at current step is small, if the safety balance in future steps is large, the algorithm should still be encouraged to explore. To capture such a new special impact from future steps, we add the term $\frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c)}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c)}$, such that $\epsilon_{h,2}$ increases with the ratio between future safety balance $\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c)$ and current balance $\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c)$. Please see Appendix B for details.

Idea III: Encouraging to explore the future subsubgraphs with higher uncertainty (i.e., looking ahead). Idea II by itself is not sufficient to achieve a sublinear regret. This is because future uncertainty could prevent the algorithm from choosing the optimal action at current step. Consider Figure 1 as an example and assume G_1 is the optimal subgraph. Even though the safety value at $h = 1$ has been precisely known, the algorithm may still not choose the optimal action $a = 1$ due to future uncertainty, e.g., it is uncertain whether the first two states at $h = 2$ are safe or not. This is another critical difference compared with the case without instantaneous constraints or with only unsafe actions. Hence, at each step, the algorithm should be encouraged to explore the state that induces a future subsubgraph with *higher* uncertainty. To achieve this, we construct a new bonus term (the term with $\epsilon_{h,3}$ in (13)) that is the maximum UCB bonus over all future subsubgraphs $G_h(s)$, where

$$
\epsilon_{h,3} = \frac{4\beta H / \tilde{\delta}}{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - \kappa}. \qquad (15)
$$

Differently from $\epsilon_{h,2}$ in (14), the term $\frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c)}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c)}$ does not appear in $\epsilon_{h,3}$, because the maximization in this bonus term is taken over all states and actions in $\mathcal{G}_h(s)$, which already captures the impacts from future steps.

Idea IV: Encouraging to explore the past subsubgraphs with higher uncertainty (i.e., peeking backward). Surprisingly, with Ideas II and III alone, a sublinear regret may still not be achieved. This is because of the tricky impact from past steps. Intuitively, by choosing a different action at step $h = 1$, what will happen in future steps could be completely different. To resolve this new challenge, we construct a new bonus term, i.e., the term with ϵ_4 in (13), to encourage LSVI-NEW to explore the past subsubgraphs with *higher* uncertainty, where

$$
\epsilon_4 = \frac{4\beta H}{\bar{c} - c_1^0 - \Delta_{\boldsymbol{\phi}}(c)}. \qquad (16)
$$

Differently from $\epsilon_{h,3}$ in (15), the denominator here depends on c_1^0 (not $\bar{c}_{h'}^0$) at step $h = 1$ that affects all future steps.

Finally, in Step-5, LSVI-NEW chooses an action

$$
a_h^k = \arg\max_{a \in \mathcal{A}_h^{k,\text{safe}}(s,a)} Q_h^k(s_h^k, a). \qquad (17)
$$

4. Theoretical Results

In this section, we provide the safety and regret guarantees for our LSVI-NEW algorithm, and a regret lower-bound.

Before these, we make two assumptions for obtaining good theoretical performance in our setting. Assumption 3 below is from Amani et al. (2021). We let $\Phi_{\boldsymbol{\alpha}}(s,a) \triangleq [\alpha(s')\boldsymbol{\phi}(s,a,s')]_{s'\in\mathcal{S}(s,a)}$ denote a matrix with $\alpha(s')\boldsymbol{\phi}(s,a,s')$ in each column, where $\alpha(s')$ is a scalar.

Assumption 3. (Star convexity) For all states s_h at step h, the set $\mathcal{D}(s_h) \triangleq \{\Phi_{\mathbf{1}}(s_h,a) : a \in \mathcal{A}\} \cup \{\Phi_{\mathbf{1}}(s_h^0,a_h^0) : \Phi_{\mathbf{1}}(s_h^0,a_h^0,\cdot) = \boldsymbol{\phi}(s_h^0,a_h^0,s_{h+1}^0)\}$ is a star convex set around the safe feature $\boldsymbol{\phi}(s_h^0,a_h^0,s_{h+1}^0)$, i.e., for all $\Phi_{\mathbf{1}}(s_h,a) \in \mathcal{D}(s_h)$ and $\boldsymbol{\alpha} : \mathcal{S}_h(s_h,a) \rightarrow [0,1]$ with $\|\boldsymbol{\alpha}\|_1 = 1$, we have $\Phi_{\boldsymbol{\alpha}}(s_h,a) + \Phi_{\mathbf{1}-\boldsymbol{\alpha}}(s_h^0,a_h^0) \in \mathcal{D}(s_h)$, where $\mathbf{1}$ denotes a vector with all entries equal to 1.

Next, we let $f_h(\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2) \triangleq \frac{\|\boldsymbol{\phi}_1-\boldsymbol{\phi}_2\|_2}{\|\boldsymbol{\phi}(s_h^*,a_h^*,s_{h+1}^*)-\boldsymbol{\phi}(s_h^0,a_h^0,s_{h+1}^0)\|_2}$ denote the $\mathcal{L}_2$-distance between features $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$, normalized by the $\mathcal{L}_2$-distance between the unknown optimal feature $\boldsymbol{\phi}(s_h^*,a_h^*,s_{h+1}^*)$ and the known safe feature $\boldsymbol{\phi}(s_h^0,a_h^0,s_{h+1}^0)$ at step h. Let $g(r_{h,1}-r_{h,2}) \triangleq \frac{|r_{h,1}-r_{h,2}|}{r_h(s_h^*,a_h^*)}$ denote reward difference $|r_{h,1}-r_{h,2}|$, normalized by the reward of the unknown optimal state-action pair at step h.

Assumption 4. (Lipschitz rewards and transitions) There exists $\delta \in (0,1)$, s.t., for any two safe state-action pairs $(s(i),a(i))$ and $(s(j),a(j))$ at step h,

$$\begin{aligned} & g\left(r_h(s(i),a(i)) - r_h(s(j),a(j))\right) \\ & \qquad \leq \delta f_h\left(\boldsymbol{\phi}(s(i),a(i),\cdot) - \boldsymbol{\phi}(s(j),a(j),\cdot)\right), \quad (18) \\ & f_{h'}\left(\boldsymbol{\phi}(s_{h'}(i),a_{h'}(i),\cdot) - \boldsymbol{\phi}(s_{h'}(j),a_{h'}(j),\cdot)\right) \\ & \qquad \leq \delta f_h\left(\boldsymbol{\phi}(s(i),a(i),\cdot) - \boldsymbol{\phi}(s(j),a(j),\cdot)\right), \quad (19) \end{aligned}$$

where $(s_{h'}(i),a_{h'}(i))$ $(h'>h)$ is the descendant of the state-action pair $(s(i),a(i))$ in the safe subgraphs.

Note that (18) implies that rewards are δ-Lipschitz: as feature differences (RHS of (18)) become smaller, reward differences (LHS of (18)) become smaller; and (19) implies that safe transitions are δ-Lipschitz: as feature differences at current step (RHS of (19)) become smaller, feature differences at future steps h' (LHS of (19)) become smaller.

When the unsafe states and transitions are taken into consideration, to still achieve a sublinear regret, Assumption 4 is required. This is because (i) if rewards are not Lipschitz, even though a feature vector *close to* the optimal one is learned to be safe, the learner could still suffer from a large reward gap compared with the optimal safe decision, which could result in a linear-in-T regret; (ii) if safe transitions are not Lipschitz, even though the optimal safe decision at a step *has been learned*, the learner could still be far away from optimum *in future steps*, and hence suffer from a large reward gap, which could also result in a linear-in-T regret.

4.1. Performance Guarantees and A Lower Bound

In Theorem 1 below, we show that LSVI-NEW is safe.

Theorem 1. ***(Safety)*** *For any $p \in (0,1)$, with probability $1-p$, our LSVI-NEW algorithm satisfies the instantaneous hard constraint (1) at all steps h of all episodes k.*

Thanks to our Idea I in Section 3 for guaranteeing safety, the proof of Theorem 1 (in Appendix A) focuses on quantifying the accuracy of the estimated safety value in (8). Below, Theorem 2 provides the regret upper-bound of LSVI-NEW.

Theorem 2. ***(Regret)*** *By setting $\tilde{\delta} = \delta$, $\lambda = d$, $\beta = \max\left\{\sigma\sqrt{d\log\left(\frac{2+2TD^2/\lambda}{p}\right)} + \sqrt{\lambda}L, b_\beta dH\sqrt{\log\left(\frac{dT}{p}\right)}\right\}$, $K' = 4\beta D\sqrt{T}\log\left(\frac{d}{p}\right)$, where $T = HK$, $\kappa = \frac{4\beta D}{\lambda+\lambda_0 K'}$ and $\Delta_c = \bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c)$, then there exist absolute constants $b_\beta > 0$ and $\lambda_0 > 0$, with probability $1-p$, the regret of LSVI-NEW is upper-bounded by*

$$\begin{aligned} & \left[\epsilon_1 + \epsilon_4 + \max_h\left(\epsilon_{h,2} + \epsilon_{h,3}\right)\right] \cdot \sqrt{2dHT\log(1+T)} \\ & + 2H\sqrt{T\log\left(\frac{2dT}{p}\right)} + HK' + \frac{D}{\lambda_0}\left(\frac{K}{K'} - 1\right). \quad (20) \end{aligned}$$

The regret in (20) is dominated by the first term (the first line in (20)) that results from the aforementioned new challenges due to the instantaneous hard constraint. Thus, incorporated with the values of the parameters, Theorem 2 indicates that the regret of LSVI-NEW is upper-bounded by $\tilde{O}\left(\frac{dH^3\sqrt{dK}}{\bar{c}-\bar{c}_1^0-\Delta_{\boldsymbol{\phi}}(c)}\right)$. Notably, it nearly matches the state-of-the-art regret $\tilde{O}\left(\frac{dH^3\sqrt{dK}}{\bar{c}-\bar{c}_1^0-\Delta_{\boldsymbol{\phi}}(c)}\right)$ in the setting with only unsafe actions (Amani et al., 2021) and $\tilde{O}(dH^2\sqrt{K})$ in the unconstrained linear mixture MDP (Jia et al., 2020), while comparing with different optimal policies. *To the best of our knowledge, this is the first such result in the literature.* Further, we provide a lower bound in Theorem 3 below.

Theorem 3. ***(A lower bound)*** *Assuming $K \geq 32\underline{R}$. The regret of any safe algorithm π is lower-bounded as follows:*

$$R^\pi \geq \underline{R} \triangleq \max\left\{\frac{dH\sqrt{K}}{16\sqrt{2}}, \frac{H/24}{(\bar{c}-\bar{c}_1^0-\Delta_{\boldsymbol{\phi}}(c))^2}\right\}. \quad (21)$$

Theorem 3 implies that the dependency of the regret of LSVI-NEW on $\bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c)$ is necessary. In addition, the regret of LSVI-NEW matches the lower bound within a factor of $\tilde{O}(H^2\sqrt{d})$. Same as in the setting with only unsafe actions, we conjecture that this gap can be further reduced by applying Bernstein inequality and leave this as future work. Please see Appendix F for the proof.

4.2. Proof Sketch for Theorem 2

In this subsection, we provide the high-level ideas for proving Theorem 2 (please see Appendix E for the proof). Because of the new challenges from instantaneous hard constraints and our novel ideas in the algorithm design, there are several new difficulties in the regret analysis. The key ones are: (I) Differently from MDPs without constraints or with only unsafe actions, in our case, different policies could visit very different sets of states at each step. Hence, the commonly-used invariant on V-values that relies on the *ergodicity* property no longer holds. (II) How to quantify the impacts when looking ahead and peeking backward. Below, we introduce our new analytical ideas, which may be of independent interest.

Step-I: Solving difficulty I by constructing new invariants. We construct new forms of V-value functions for different policies below. We let $\mathcal{S}_h^*$ denote the state set at step h in the optimal safe subgraph. Let $\mathcal{S}_h^k$ denote the state set at step h in the subgraph followed by policy π^k of LSVI-NEW in episode k. Moreover, we let $\tilde{f}_h(s,a) \triangleq f_h(\boldsymbol{\phi}(s,a,\cdot) - \boldsymbol{\phi}(s_h^*, a_h^*, \cdot))$ denote the gap of transitions compared with optimal transitions. Let $\tilde{\mathcal{A}}_h^k(s) \triangleq \{a \in \mathcal{A}_h^{k,\text{safe}}(s) : \tilde{f}_h(s,a) \leq \bar{\alpha}_0\} \cup \{a_h^k(s)\}$ capture the safe actions with transitions close to the optimal transitions, where $\bar{\alpha}_0$ is the maximum of α_0 in (28) and the RHS of (29). Let $\tilde{\mathcal{S}}_h^k \triangleq \{s \in \mathcal{S}_h^{k,\text{safe}} : \exists a \in \mathcal{A}_h^{k,\text{safe}}(s), \text{ s.t., } \tilde{f}_h(s,a) \leq \bar{\alpha}_0\} \cup \mathcal{S}_h^k$ capture the safe states with transitions close to the optimal transitions. Next, we define the V-value functions of the optimal policy, estimated policy and policy π^k to be

$$V_h^*(s) \triangleq Q_h^*(s, a_h^*(s)), \forall s \in \mathcal{S}_h^*, \tag{22}$$

$$V_h^k(s) \triangleq \max_{a \in \tilde{\mathcal{A}}_h^k(s)} Q_h^k(s,a), \forall s \in \tilde{\mathcal{S}}_h^k, \tag{23}$$

$$V_h^{\pi^k}(s) \triangleq Q_h^{\pi^k}(s, a_h^k(s)), \forall s \in \mathcal{S}_h^k, \tag{24}$$

respectively. Then, the regret $R^{\text{LSVI-NEW}}$ can be decomposed into two parts, i.e., the values in the two brackets $[\cdot]$ below:

$$\sum_{k=1}^{K} \left\{ [V_1^*(s_1) - V_1^k(s_1)] + [V_1^k(s_1) - V_1^{\pi^k}(s_1)] \right\}. \tag{25}$$

To upper-bound the regret, we prove that, with high probability, (i) the value in the first bracket of (25) is non-positive; (ii) the value in the second bracket can be upper-bounded. Result (ii) can be obtained by upper-bounding the bonus terms, which can further be proven by slightly modifying existing techniques in linear mixture MDP. The main difficulty is to prove result (i). To resolve this difficulty, we construct two new invariants that hold at each step.

Lemma 1. ***(New invariants)*** *At each step h of each episode, (i) for any state s, s.t., $s \in \mathcal{S}_h^*$ and $s \in \tilde{\mathcal{S}}_h^k$, we have*

$$V_h^k(s) \geq V_h^*(s); \tag{26}$$

(ii) for any state s, s.t., $s \in \mathcal{S}_h^*$ and $s \notin \tilde{\mathcal{S}}_h^k$, and any state $\hat{s}$, s.t., $\hat{s} \in \tilde{\mathcal{S}}_h^k$ and $\hat{s} \notin \mathcal{S}_h^*$, we have

$$V_h^k(\hat{s}) \geq V_h^*(s). \tag{27}$$

Invariant (i) shows that, if the optimal state has been found, the estimated V-value must be higher than the optimal V-value. Notice that if the optimal safe action has also been found, (26) trivially holds. If it has not been found, thanks to our new bonus terms that essentially capture the distance from the optimal action, (26) still holds. Moreover, invariant (ii) shows that, if the optimal state has not been found, the V-value of the sub-optimal state in $\tilde{\mathcal{S}}_h^k$ is still larger than the optimal V-value. This is intuitively because $\tilde{\mathcal{S}}_h^k$ only contains safe states with transitions *close* to the optimal transitions, and the distance is captured by our new bonus terms. Please see Appendix D for details and the proof.

Step-II: Solving difficulty II by quantifying future impacts. The impact when looking ahead can be characterized by quantifying the impacts from future steps.

Lemma 2. ***(Impacts from future steps)*** *For any state s, s.t., $s \in \mathcal{S}_h^*$ and $s \in \tilde{\mathcal{S}}_h^k$, if $a_h^*(s) \notin \tilde{\mathcal{A}}_h^k(s)$, there must exist an action $a_0 \in \tilde{\mathcal{A}}_h^k(s)$, s.t.,*

$$\tilde{f}_h(s, a_0 | s_h^* = s) \leq \alpha_0, \tag{28}$$

where $\alpha_0 = 1 - \frac{(\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) - l_1)(\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - l_2)}{(\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) + l_1)(\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + l_2)}$, $l_1 = 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}$ *and* $l_2 = 2\beta \max_{\{h < h' \leq H, (s_{h'}^*, a_{h'}^*), s'\}} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}$.

Lemma 2 implies that when k increases, the UCB terms l_1 and l_2 decrease to be closer to 0, and thus α_0 gets closer to 0. Then, the gap between LSVI-NEW's decision and the optimal decision, i.e., $\tilde{f}_h(s, a_0 | s_h^* = s)$ on the LHS of (28), gets closer to 0. This is consistent with the intuition that as more safety values revealed, we should be able to get closer to the optimal action. Moreover, when there is no constraint on states, all terms related to the next state s' in α_0 would be 0. Then, α_0 would be reduced to be $1 - \frac{\bar{c} - c_h^0 - 2\beta \|\boldsymbol{\phi}(s, a_h^*(s))\|}{\bar{c} - c_h^0}$, which results in a parameter same to that used in the case with only unsafe actions (Amani et al., 2021). However, due to unsafe states and transitions, impacts from future steps $h' > h$ are captured in α_0 here, which results in a different parameter $\epsilon_{h,2}$ in our Idea II and a new parameter $\epsilon_{h,3}$ in Idea III. Please see Appendix B for details and the proof.

Step-III: Solving difficulty II by quantifying past impacts. The impact when peeking backward can be characterized by quantifying the impacts from past steps.

Lemma 3. ***(Impacts from past steps)*** *For any state $\hat{s}$, s.t., $\hat{s} \in \tilde{\mathcal{S}}_h^k$ and $\hat{s} \notin \mathcal{S}_h^*$, there must exist an action $a_0 \in \tilde{\mathcal{A}}_h^k(\hat{s})$ and $1 \leq h' \leq h$, s.t.,*

$$\tilde{f}_h(\hat{s}, a_0) \leq 1 - \frac{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - l_3}{\delta(\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + l_3)}, \tag{29}$$

where $l_3 = 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}$.

Table 1. Comparison of the average rewards (RW) and constraint violations (CV) in synthetic environments.

	LSVI-NEW	UCRL	SLUCB-QVI	sMDP
RW	1.53	1.78	1.59	0.85
CV	0	0.91	0.83	0

Differently from Lemma 2, Lemma 3 quantifies the impacts from past steps, i.e., $h' \leq h$. This special impact results in the new bonus term with parameter ϵ_4 in our Idea IV in Section 3. Moreover, Lemma 3 implies that when the number of episodes k increases, the UCB term l_3 decreases to be closer to 0, and thus the RHS of (29) in Lemma 3 gets closer to 0. This implies that $\tilde{f}_h(\hat{s}, a_0)$ on the LHS of equation (29) gets closer to 0. Notice that $\tilde{f}_h(\hat{s}, a_0)$ represents the gap between the decision of the policy π^k used by LSVI-NEW and the optimal decision. In addition, in the RHS of equation (29), l_3 characterizes the transition uncertainty. Therefore, the above implication is consistent with the intuition that as more safety values are revealed, we should be able to get closer to the optimal action. See Appendix C for the proof.

These new difficulties are also the reasons that all $\epsilon_{h,2}$, $\epsilon_{h,3}$ and ϵ_4 are different from the parameter used in the setting with only unsafe actions (Amani et al., 2021).

5. Numerical Results

In this section, we provide some numerical results. We simulate two types of environments: (i) synthetic environments and (ii) robot path planning environments (Wachi et al., 2018). In these experiments, we compare our LSVI-NEW algorithm with the UCRL algorithm in Jia et al. (2020), the SLUCB-QVI algorithm from Amani et al. (2021) and the sMDP algorithm in Wachi et al. (2018). For all algorithms, we compare their average (averaged over the number K of episodes) rewards (RW) and number of constraint violations (CV), in terms of the average number of episodes that the instantaneous hard constraint is violated.

(i) Synthetic environments: To generate the environments, we set the dimension $d = 5$, length of each episode $H = 3$, number of episodes $K = 10000$, and the safety threshold $\bar{c} = 0.5$. The transition-probability parameter $\boldsymbol{\mu}_h^*$ and safety-value parameter $\boldsymbol{\gamma}_h^*$ are generated according to the truncated multivariate Gaussian distribution $\mathcal{N}(0, \mathbf{1}_d)$. From Table 1, we can see that although UCRL and SLUCB-QVI (which disregard the instantaneous hard constraint) obtain higher reward, they violate the constraint for most episodes. In contrast, our LSVI-NEW algorithm satisfies the constraint all the time. On the other hand, although the sMDP algorithm does not violate the constraint, its reward is low since it is too conservative. In contrast, our LSVI-NEW algorithm achieves a larger reward in our simulated environments.

Table 2. Comparison of the average rewards (RW) and constraint violations (CV) in robot path planning environments.

	LSVI-NEW	UCRL	SLUCB-QVI	sMDP
RW	53.6	59.6	54.9	33.1
CV	0	0.81	0.71	0

(ii) Robot path planning environments: We consider a 10×10 2D-map. The state represents the location of the robot in the map. The action at each location represents the degree of the robot's searching direction. The unsafe states correspond to the location where there exists a rock or a cliff. Differently from Wachi et al. (2018) where a deterministic transition is assumed, we consider the uncertainty between the real motion and the command. That is, by taking an action with a degree of x, the robot could either move directly in the targeting direction or deviate from the targeting direction by a certain degree. Moreover, we change the episode length to be $H = 100$. Similarly to the conclusion in synthetic environments, from Table 2, we can see that LSVI-NEW is better in satisfying the instantaneous hard constraint and obtaining a higher reward in our simulated environments.

6. Conclusion

In this paper, we make the first effort to resolve the challenges due to unsafe states and actions under instantaneous hard constraints in RL. We develop an RL algorithm that achieves a regret that nearly matches the state-of-the-art regret in the setting with only unsafe actions and that in the unconstrained setting (while comparing with different optimal policies), and is safe (i.e., satisfies the instantaneous hard constraint) at each step. We also provide a lower bound of the regret that indicates that the dependency of the regret of our algorithm on the safety parameters is necessary. Further, both our algorithm design and regret analysis involve several novel ideas, which may be of independent interest. An interesting future work is to study the impact of number N of instantaneous hard constraints on the regret. Note that the construction for the bonus terms will need to be more careful, e.g., the bonus terms need to increase with $\log N$. Thus, we conjecture that the final regret will depend on $\log N$, which would be similar to that in the setting with soft cumulative constraints (HasanzadeZonuzy et al., 2021).

Acknowledgements

This work has been partly supported by NSF grants NSF AI Institute (AI-EDGE) CNS-2112471, RINGS-2148253, CNS-2106933, 2007231, CNS-1955535, and CNS-1901057, and by Army Research Office under Grant W911NF-21-1-0244.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. PMLR, 2017.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pp. 10–4, 2019.

Amani, S., Alizadeh, M., and Thrampoulidis, C. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

Badiei, M., Li, N., and Wierman, A. Online convex optimization with ramp constraints. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 6730–6736. IEEE, 2015.

Bai, Q., Bedi, A. S., Agarwal, M., Koppel, A., and Aggarwal, V. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3682–3689, 2022.

Brantley, K., Dudik, M., Lykouris, T., Miryoosefi, S., Simchowitz, M., Slivkins, A., and Sun, W. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *Advances in Neural Information Processing Systems*, 33:16315–16326, 2020.

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

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

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

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

HasanzadeZonuzy, A., Bura, A., Kalathil, D., and Shakkottai, S. Learning with safety constraints: Sample complexity of reinforcement learning for constrained mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7667–7674, 2021.

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

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pp. 666–686. PMLR, 2020.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Kalagarla, K. C., Jain, R., and Nuzzo, P. A sample-efficient algorithm for episodic finite-horizon mdp with constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8030–8037, 2021.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17 (1):1–42, 2016.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, Y., Qu, G., and Li, N. Online optimization with predictions and switching costs: Fast algorithms and the fundamental limit. *IEEE Transactions on Automatic Control*, 66(10):4761–4768, 2020.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., and Tian, C. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021.

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

Paternain, S., Calvo-Fullana, M., Chamon, L. F., and Ribeiro, A. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 2022.

Shi, M., Lin, X., and Fahmy, S. Competitive online convex optimization with switching costs and ramp constraints. *IEEE/ACM Transactions on Networking*, 29(2):876–889, 2021a.

Shi, M., Lin, X., and Jiao, L. Combining regularization with look-ahead for competitive online convex optimization. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pp. 1–10. IEEE, 2021b.

Shi, M., Lin, X., and Jiao, L. Power-of-2-arms for bandit learning with switching costs. In *Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 131–140, 2022a.

Shi, Y., Qu, G., Low, S., Anandkumar, A., and Wierman, A. Stability constrained reinforcement learning for real-time voltage control. In *2022 American Control Conference (ACC)*, pp. 2715–2721. IEEE, 2022b.

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

Turchetta, M., Berkenkamp, F., and Krause, A. Safe exploration in finite markov decision processes with gaussian processes. *Advances in Neural Information Processing Systems*, 29, 2016.

Vamvoudakis, K. G., Wan, Y., Lewis, F. L., and Cansever, D. *Handbook of Reinforcement Learning and Control*. Springer, 2021.

Wachi, A., Sui, Y., Yue, Y., and Ono, M. Safe exploration and optimization of constrained mdps using gaussian processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Wei, H., Liu, X., and Ying, L. A provably-efficient model-free algorithm for constrained markov decision processes. *arXiv preprint arXiv:2106.01577*, 2021.

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

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

Yang, T.-Y., Rosca, J., Narasimhan, K., and Ramadge, P. J. Projection-based constrained policy optimization. In *International Conference on Learning Representations*, 2019.

Zhou, D. and Gu, Q. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *Advances in Neural Information Processing Systems*, 35, 2022.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021a.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021b.

A. Proof of Theorem 1

Remember that our Idea I in Section 3 is mainly designed for guaranteeing safety. As we discussed there, (i) condition 1 in (9) implies that by choosing action a for state s at step h, the instantaneous hard constraint is guaranteed to be satisfied at step h; (ii) condition 2 in (10) implies that all possible next states in $\mathcal{S}_h(s,a)$ (i.e., the next states that could be visited with non-zero probability) must be safe for next step $h+1$. Thus, with conditions 1 and 2 satisfied simultaneously in a backward manner, all step $h' \geq h$ (not even just next step $h+1$) following (s,a) must be safe. Hence, the probability of our LSVI-NEW algorithm being safe depends on the accuracy of the estimated safety value $\tilde{c}_h^k$ in (8).

Moreover, remember that, on the RHS of (8), the first term is the projected safety value of (s,a,s') on $\mathcal{U}_h$, the second term is the projected empirical safety value of (s,a,s') on $\mathcal{U}_h^{\perp}$, and the last term is a UCB bonus for the safety uncertainty. In addition, the second term there relies on the accuracy of the regularized least-square estimator of the projected safety parameter $\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\gamma}_h^*)$. Thus, the accuracy of $\tilde{c}_h^k$ further depends on how accurate $\boldsymbol{\gamma}_h^k$ in (7) is and how small the safety uncertainty is.

Therefore, we first prove Lemma 4 below for quantifying the accuracy of the estimated safety parameter $\boldsymbol{\gamma}_h^k$ in (7).

Lemma 4. ***(Accuracy of the estimated safety parameter)*** *For any $p \in (0,1)$, with probability $1-p$, we have that, for all steps h of all episode k,*

$$\left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\gamma}_h^*) - \boldsymbol{\gamma}_h^k\right\|_{\boldsymbol{\Lambda}_{h,1}^k} \leq \beta_1, \tag{30}$$

where $\beta_1 = \sigma\sqrt{d \log\left(\frac{2+\frac{2TD^2}{\lambda}}{p}\right)} + \sqrt{\lambda}L.$

Proof. **(Proof of Lemma 4)** First, according to (7), we have that the estimated safety parameter is equal to

$$\begin{aligned}
\boldsymbol{\gamma}_h^k = &\left(\lambda \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I}) + \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \boldsymbol{\psi}^{\mathrm{T}}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau})\right)^{-1} \\
&\qquad \cdot \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \left(\hat{c}_h^{\tau} - \frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}_{h,h+1}^{\tau}), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0\right) \\
= &\left(\lambda \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I}) + \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \boldsymbol{\psi}^{\mathrm{T}}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau})\right)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \left[\langle \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\gamma}_h^*), \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau})\rangle + \zeta_h^{\tau}\right].
\end{aligned}$$

By opening the bracket $[\cdot]$, and adding and subtracting the term $\lambda\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I})$, we have

$$\begin{aligned}
\boldsymbol{\gamma}_h^k = &\left(\lambda \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I}) + \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \boldsymbol{\psi}^{\mathrm{T}}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau})\right)^{-1} \left(\lambda \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I}) + \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \boldsymbol{\psi}^{\mathrm{T}}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau})\right) \\
&\cdot \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\gamma}_h^*) - \left(\lambda \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I}) + \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \boldsymbol{\psi}^{\mathrm{T}}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau})\right)^{-1} \lambda \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I}) \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\gamma}_h^*) \\
&+ \left(\lambda \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I}) + \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \boldsymbol{\psi}^{\mathrm{T}}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau})\right)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \zeta_h^{\tau}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\boldsymbol{\gamma}_h^k = \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\gamma}_h^*) &- \left(\lambda \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I}) + \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \boldsymbol{\psi}^{\mathrm{T}}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau})\right)^{-1} \lambda \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I}) \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\gamma}_h^*) \\
&+ \left(\lambda \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{I}) + \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \boldsymbol{\psi}^{\mathrm{T}}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau})\right)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}_{h,h+1}^{\tau}) \zeta_h^{\tau}.
\end{aligned} \tag{31}$$

According to (31), the square of the left-hand-side of (30) is equal to

$$\left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*)-\boldsymbol{\gamma}_h^k\right\|^2_{\boldsymbol{\Lambda}_{h,1}^k} = \left[\left(\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*)-\boldsymbol{\gamma}_h^k\right)\boldsymbol{\Lambda}_{h,1}^k\right]^{\mathrm{T}}\left(\lambda\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{I})+\sum_{\tau=1}^{k-1}\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\phi}_{h,h+1}^{\tau})\boldsymbol{\psi}^{\mathrm{T}}(\mathcal{U}_h^{\perp},\boldsymbol{\phi}_{h,h+1}^{\tau})\right)^{-1}$$
$$\cdot\left(\lambda\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{I})\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*)-\sum_{\tau=1}^{k-1}\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\phi}_{h,h+1}^{\tau})\zeta_h^{\tau}\right).$$

Then, according to the Cauchy-Schwarz inequality, we have

$$\left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*)-\boldsymbol{\gamma}_h^k\right\|^2_{\boldsymbol{\Lambda}_{h,1}^k} \leq \left\|\left(\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*)-\boldsymbol{\gamma}_h^k\right)\boldsymbol{\Lambda}_{h,1}^k\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}$$
$$\cdot\left[\left\|\lambda\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{I})\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*)\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}+\left\|\sum_{\tau=1}^{k-1}\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\phi}_{h,h+1}^{\tau})\zeta_h^{\tau}\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}\right].$$

Notice that the smallest eigenvalue of $\boldsymbol{\Lambda}_{h,1}^k$ is $\lambda_{min}(\boldsymbol{\Lambda}_{h,1}^k)=\lambda$. Hence, according to Theorem 1 in Abbasi-Yadkori et al. (2011), we have that, with probability $1-p$ for any $p\in(0,1)$,

$$\left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*)-\boldsymbol{\gamma}_h^k\right\|^2_{\boldsymbol{\Lambda}_{h,1}^k} \leq \left\|\left(\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*)-\boldsymbol{\gamma}_h^k\right)\boldsymbol{\Lambda}_{h,1}^k\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}\cdot\left[\sigma\sqrt{d\log\left(\frac{2+\frac{2kD^2}{\lambda}}{p}\right)}+\sqrt{\lambda}L\right]. \tag{32}$$

Finally, by rearranging the terms in (32), we have

$$\left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*)-\boldsymbol{\gamma}_h^k\right\|_{\boldsymbol{\Lambda}_{h,1}^k}\leq\sigma\sqrt{d\log\left(\frac{2+\frac{2kD^2}{\lambda}}{p}\right)}+\sqrt{\lambda}L\leq\beta_1.$$

This concludes the proof of Lemma 4.

□

Lemma 4 shows that with high probability, the estimated safety parameter $\boldsymbol{\gamma}_h^k$ is close enough to the projected true safety parameter $\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*)$. Now, we prove Theorem 1 based on our Idea I in Section 3 and Lemma 4 above.

Proof. **(Proof of Theorem 1)** We let $\mathcal{G}_h^{k,\text{safe}}$ denote the set of safe subsubgraphs constructed at step h in episode k by LSVI-NEW using our Idea I. Then, using mathematical induction, we prove that $\mathcal{G}_h^{k,\text{safe}}$ is safe, i.e., any state-action-state triplet $(s_{h'}^k,a_{h'}^k,s_{h'+1}^k)$, where $h\leq h'\leq H$, in $\mathcal{G}_h^{k,\text{safe}}$ satisfies the instantaneous hard constraint (1).

(i) Base case: when $h=H$, according to Lemma 4 and the Cauchy-Schwarz inequality, we have

$$\left\langle\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*)-\boldsymbol{\gamma}_h^k,\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\phi}(s_H^k))\right\rangle\leq\beta_1\left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\phi}(s_H^k))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}. \tag{33}$$

From (33), we have

$$\left\langle\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*),\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\phi}(s_H^k))\right\rangle\leq\left\langle\boldsymbol{\gamma}_h^k,\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\phi}(s_H^k))\right\rangle+\beta_1\left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\phi}(s_H^k))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}. \tag{34}$$

Next, since the left-hand-side of (34) is equal to

$$\left\langle\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\gamma}_h^*),\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\phi}(s_H^k))\right\rangle=\left\langle\boldsymbol{\gamma}_h^*,\boldsymbol{\phi}(s_H^k)\right\rangle-\left\langle\boldsymbol{\gamma}_h^*,\boldsymbol{\psi}(\mathcal{U}_h,\boldsymbol{\phi}(s_H^k))\right\rangle$$
$$=\left\langle\boldsymbol{\gamma}_h^*,\boldsymbol{\phi}(s_H^k)\right\rangle-\frac{\langle\boldsymbol{\psi}(\mathcal{U}_h,\boldsymbol{\phi}(s_H^k)),\tilde{\boldsymbol{\phi}}(s_H^0)\rangle}{\|\boldsymbol{\phi}(s_H^0)\|_2}\cdot c_H^0,$$

we have

$$\left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s_H^k) \right\rangle \leq \frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s_H^k)), \tilde{\boldsymbol{\phi}}(s_H^0) \rangle}{\|\boldsymbol{\phi}(s_H^0)\|_2} \cdot c_H^0 + \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s_H^k) \right\rangle + \beta_1 \left\| \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s_H^k)) \right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}. \tag{35}$$

Notice that, since the parameter β used for the estimated safety value $\tilde{c}_H^k(s_H^k)$ in (8) is larger than or equal to β_1, the right-hand-side of (35) is less than or equal to $\tilde{c}_H^k(s_H^k)$, which is less than or equal to $\bar{c}$ due to our condition 1 in (9). Hence, we have $c_H(s_H^k) \leq \bar{c}$.

(ii) Induction step: we hypothesize that $\mathcal{G}_h^{k,\text{safe}}$ is safe when $h = h_0$. Then, we prove that $\mathcal{G}_h^{k,\text{safe}}$ is safe for $h = h_0 - 1$ similar to the base case, while condition 2 that we construct in (10) becomes important here. First, according to Lemma 4 and the Cauchy-Schwarz inequality, we have

$$\left\langle \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\gamma}_h^*) - \boldsymbol{\gamma}_h^k, \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k)) \right\rangle \leq \beta_1 \left\| \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k)) \right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}. \tag{36}$$

From (36), we have

$$\left\langle \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\gamma}_h^*), \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k)) \right\rangle \leq \left\langle \boldsymbol{\gamma}_h^k, \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k)) \right\rangle + \beta_1 \left\| \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k)) \right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}. \tag{37}$$

Next, since the left-hand-side of (37) is equal to

$$\begin{aligned} \left\langle \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\gamma}_h^*), \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k)) \right\rangle &= \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k) \right\rangle - \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k)) \right\rangle \\ &= \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k) \right\rangle - \frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k)), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0) \rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0, \end{aligned}$$

we have

$$\begin{aligned} &\left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k) \right\rangle \\ &\leq \frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k)), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0) \rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0 + \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k) \right\rangle + \beta_1 \left\| \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s_h^k, a_h^k, s_{h+1}^k)) \right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}. \end{aligned} \tag{38}$$

Notice that, the right-hand-side of (38) is less than or equal to the estimated safety value $\tilde{c}_h^k(s_h^k, a_h^k, s_{h+1}^k)$ in (8), which is less than or equal to $\bar{c}$ due to our condition 1 in (9). Thus, we have $c_h(s_h^k) \leq \bar{c}$. In addition, according to condition 2 that we construct in (10) and the induction hypothesis, s_{h+1} must also be safe. Hence, $\mathcal{G}_h^{k,\text{safe}}$ is safe.

□

B. Proof of Lemma 2

As we discussed in Section 4.2, Lemma 2 implies that when k increases, the UCB terms l_1 and l_2 decrease to be closer to 0, and thus α_0 on the right-side of (28) gets closer to 0. Then, $\tilde{f}_h(s, a_0 | s_h^* = s)$ on the left-hand-side of (28) gets closer to 0. Notice that $\tilde{f}_h(s, a_0 | s_h^* = s)$ represents the gap between the decision of the policy π^k used by LSVI-NEW and the optimal decision. In addition, in α_0, l_1 characterizes the transition uncertainty and l_2 characterizes the uncertainty from future steps. Thus, the above implication from Lemma 2 is consistent with the intuition that as more safety values revealed, we should be able to get closer to the optimal action.

Moreover, when there is no constraint on states, all terms related to the next state s' in α_0, e.g., l_2, $\bar{c}_{h'}^0$ and $\Delta_{\boldsymbol{\phi}}(c)$, would be 0. Then, α_0 would be reduced to be in a much simpler form $1 - \frac{\bar{c} - c_h^0 - 2\beta \|\boldsymbol{\phi}(s, a_h^*(s))\|}{\bar{c} - c_h^0}$, which results in a parameter that is same to that used for the UCB bonus term in the case with only unsafe actions (Amani et al., 2021). However, due to unsafe states and transitions in our case, the impacts from the future steps $h' > h$ are characterized in α_0 here, which results in a different parameter $\epsilon_{h,2}$ in our Idea II and a new parameter $\epsilon_{h,3}$ in our Idea III in Section 3.

Further, as stated in Lemma 2, we only need to show there exists such a safe action $a_0 \in \tilde{\mathcal{A}}_h^k(s)$. Thus, we only need to prove the existence of an estimated safe subgraph, such that this state-action pair (s, a_0) is contained. Hence, (28) does not depends on the estimation accuracy of the Q-value parameter w_h^*.

In this section, we provide the complete proof for Lemma 2. Please see Appendix D for our discussions and proofs on how the new impacts from future steps captured in α_0 affect the requirements for choosing the parameters $\epsilon_{h,2}$ and $\epsilon_{h,3}$.

To prove Lemma 2, we first provide another new lemma below, which proves to be important. We let

$$\Delta_h(s,a,s') \triangleq \max_{s'' \in \mathcal{S}_h(s,a)} \{c_h(s,a,s'') - c_h(s,a,s')\} \tag{39}$$

denote the maximum difference between the *true* safety value $c_h(s,a,s'')$ of the state-action-state triplet (s,a,s'') for any next state $s'' \in \mathcal{S}_h(s,a)$ of the state-action pair (s,a) and the true safety value $c_h(s,a,s')$ of the given state-action-state triplet (s,a,s'). Let

$$\tilde{\Delta}_h^k(s,a,s') \triangleq \max_{s'' \in \mathcal{S}_h(s,a)} \left\{\tilde{c}_h^k(s,a,s'') - \tilde{c}_h^k(s,a,s')\right\} \tag{40}$$

denote the maximum difference between the *estimated* safety value $\tilde{c}_h^k(s,a,s'')$ of the state-action-state triplet (s,a,s'') for any next state $s'' \in \mathcal{S}_h(s,a)$ of the state-action pair (s,a) and the estimated safety value $\tilde{c}_h^k(s,a,s')$ of the given state-action-state triplet (s,a,s').

Lemma 5. ***(Relating the true and estimated safety differences)*** *The estimated safety difference $\tilde{\Delta}_h^k(s,a,s')$ can be upper-bounded by the true safety difference $\Delta_h(s,a,s')$ as follows:*

$$\tilde{\Delta}_h^k(s,a,s') \leq \Delta_h(s,a,s') + 2\beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\|_{(\mathbf{\Lambda}_{h,1}^k)^{-1}}, \tag{41}$$

where $\tilde{s}'_{\max}$ is the maximizer of (40).

Proof. **(Proof of Lemma 5)** We let $s'_{\max}$ denote the maximizer of (39). Notice that $s'_{\max}$ could be different from $\tilde{s}'_{\max}$ (the maximizer of (40)). First, the true safety difference is equal to

$$\begin{aligned}\Delta_h(s,a,s') &= \max_{s'' \in \mathcal{S}_h(s,a)} \{c_h(s,a,s'') - c_h(s,a,s')\} = c_h(s,a,s'_{\max}) - c_h(s,a,s') \\ &= \langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s,a,s'_{\max})\rangle - \langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s,a,s')\rangle .\end{aligned} \tag{42}$$

Next, the estimated safety difference is equal to

$$\begin{aligned}&\tilde{\Delta}_h^k(s,a,s') = \max_{s'' \in \mathcal{S}_h(s,a)} \left\{\tilde{c}_h^k(s,a,s'') - \tilde{c}_h^k(s,a,s')\right\} \\ &= \frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max})), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0 + \left\langle \boldsymbol{\gamma}_h^k, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\rangle + \beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\|_{(\mathbf{\Lambda}_{h,1}^k)^{-1}} \\ &\quad - \frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s,a,s')), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0 - \left\langle \boldsymbol{\gamma}_h^k, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,s'))\right\rangle - \beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,s'))\right\|_{(\mathbf{\Lambda}_{h,1}^k)^{-1}} .\end{aligned} \tag{43}$$

Considering the second term, third term, and the last two terms on the right-hand-side of (43) together, we have

$$\begin{aligned}&\left\langle \boldsymbol{\gamma}_h^k, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\rangle + \beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\|_{(\mathbf{\Lambda}_{h,1}^k)^{-1}} \\ &\qquad\qquad - \left\langle \boldsymbol{\gamma}_h^k, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,s'))\right\rangle - \beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,s'))\right\|_{(\mathbf{\Lambda}_{h,1}^k)^{-1}} \\ &= \left\langle \boldsymbol{\gamma}_h^k - \boldsymbol{\gamma}_h^*, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\rangle + \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\rangle + \beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\|_{(\mathbf{\Lambda}_{h,1}^k)^{-1}} \\ &\qquad + \left\langle \boldsymbol{\gamma}_h^* - \boldsymbol{\gamma}_h^k, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,s'))\right\rangle - \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,s'))\right\rangle - \beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,s'))\right\|_{(\mathbf{\Lambda}_{h,1}^k)^{-1}} \\ &\leq \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\rangle - \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,s'))\right\rangle + 2\beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\|_{(\mathbf{\Lambda}_{h,1}^k)^{-1}},\end{aligned} \tag{44}$$

where the inequality is by applying Lemma 4 and the Cauchy-Schwarz inequality to the first term in the third line and the first term in the fourth line in (44) above, and the fact that $\beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,s'))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} \geq 0$. Next, by combining (43) and (44), we have

$$\begin{aligned}
\tilde{\Delta}_h^k(s,a,s') \leq & \frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max})), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0 + \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\rangle \\
& - \frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s,a,s')), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0 - \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,s'))\right\rangle + 2\beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} \\
\leq & \Delta_h(s,a,s') + 2\beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}},
\end{aligned}$$

where the last inequality is because of the definition of the true safety difference $\Delta_h(s,a,s')$ in (39).

□

Lemma 5 shows that the estimated safety difference is only larger than the true safety difference by a term, i.e., $2\beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a,\tilde{s}'_{\max}))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}$, that decreases to 0 as the number of learning episodes k increases. This is consistent with the intuition that, as k increases, the estimated safety difference $\tilde{\Delta}_h^k(s,a,s')$ should get closer to the true safety difference $\Delta_h(s,a,s')$. Below, based on Lemma 5, we prove Lemma 2.

Proof. **(Proof of Lemma 2)** Recall that $\tilde{f}_h(s, a_0|s_h^* = s)$ represents the gap between the decision of the policy π^k used by LSVI-NEW and the optimal decision. Thus, now we characterize the relation between the safety values based on the state-action pair (s, a_0) and the optimal state-action pair $(s, a_h^*(s))$. First, according to the definition of estimated safety value in (8) and Assumption 3, the estimated safety value of any state-action-state triplet $(s, a_0, s'(s, a_0))$ induced by the state-action pair (s, a_0) is equal to

$$\begin{aligned}
& \tilde{c}_h^k(s, a_0, s'(s,a_0)) \\
& = \frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s,a_0,s')), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0 + \left\langle \boldsymbol{\gamma}_h^k, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a_0,s'))\right\rangle + \beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s,a_0,s'))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} \\
& = \frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \alpha_{s'}\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0) + (1-\alpha_{s'})\boldsymbol{\phi}(s, a_h^*(s), s'(s,a_h^*(s)))), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0 \\
& \qquad + \left\langle \boldsymbol{\gamma}_h^k, \boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \alpha_{s'}\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0) + (1-\alpha_{s'})\boldsymbol{\phi}(s, a_h^*(s), s'(s,a_h^*(s))))\right\rangle \\
& \qquad + \beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \alpha_{s'}\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0) + (1-\alpha_{s'})\boldsymbol{\phi}(s, a_h^*(s), s'(s,a_h^*(s))))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}. \qquad (45)
\end{aligned}$$

where we drop (s, a_0) from $s'(s, a_0)$ for simplicity. Since $\boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)) = \boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)$ and $\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)) = 0$, from (45), we have

$$\begin{aligned}
\tilde{c}_h^k(s, a_0, s'(s,a_0)) = & \alpha_{s'(s,a_0)} \cdot \frac{\langle \boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0 \\
& + (1-\alpha_{s'(s,a_0)}) \cdot \left[\frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s, a_h^*(s), s'(s,a_h^*(s)))), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0 + \left\langle \boldsymbol{\gamma}_h^k, \boldsymbol{\psi}\Big(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}\big(s, a_h^*(s), s'(s,a_h^*(s))\big)\Big)\right\rangle \right. \\
& \left. \qquad + \beta \left\| \boldsymbol{\psi}\Big(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s, a_h^*(s), s'(s,a_h^*(s)))\Big) \right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} \right]. \qquad (46)
\end{aligned}$$

Let us focus on the terms in the bracket $[\cdot]$ of (46). Notice that, (i) we have

$$\begin{aligned}
&\frac{\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s, a_h^*(s), s'(s, a_h^*(s)))), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\rangle}{\|\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)\|_2} \cdot c_h^0 + \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s, a_h^*(s), s'(s, a_h^*(s))))\right\rangle \\
&= \left\langle \boldsymbol{\gamma}_h^*, \left\langle \boldsymbol{\psi}(\mathcal{U}_h, \boldsymbol{\phi}(s, a_h^*(s), s'(s, a_h^*(s)))), \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\right\rangle \tilde{\boldsymbol{\phi}}(s_h^0, a_h^0, s_{h+1}^0)\right\rangle + \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s, a_h^*(s), s'(s, a_h^*(s))))\right\rangle \\
&= \langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s, a_h^*(s), s'(s, a_h^*(s)))\rangle \\
&= c_h(s, a_h^*(s), s'(s, a_h^*(s))) \\
&\le \bar{c} - \Delta_h(s, a_h^*(s), s'(s, a_h^*(s))),
\end{aligned} \tag{47}$$

where the inequality is (a) because $(s, a_h^*(s))$ is safe, and hence $c_h(s, a_h^*(s), s') \le \bar{c}$ for all $s' \in \mathcal{S}_h(s, a_h^*(s))$; (b) according to the definition of the true safety difference in (39). (ii) According to Lemma 4, we have

$$\left\langle \boldsymbol{\gamma}_h^k - \boldsymbol{\gamma}_h^*, \boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s, a_h^*(s), s'(s, a_h^*(s))))\right\rangle \le \beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s, a_h^*(s), s'(s, a_h^*(s))))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}. \tag{48}$$

By combining (46), (47) and (48), we have

$$\begin{aligned}
&\tilde{c}_h^k(s, a_0, s'(s, a_0)) \\
&\le \alpha_{s'(s,a_0)} c_h^0 + (1 - \alpha_{s'(s,a_0)}) \left[\bar{c} - \Delta_h(s, a_h^*(s), s'(s, a_h^*(s))) + 2\beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s, a_h^*(s), s'(s, a_h^*(s))))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}\right].
\end{aligned} \tag{49}$$

Next, since the optimal action $a_h^*(s)$ has not been found by the algorithm, there must exist at least one next-state $s' \in \mathcal{S}_h(s, a)$, such that the instantaneous hard constraint (1) is violated. Thus, we must have

$$\tilde{c}_h^k(s, a_h^*(s), \tilde{s}'_{\max}) > \bar{c}. \tag{50}$$

Combining (50) and Lemma 5, we have that, for all next state $s'(s, a_h^*(s)) \in \mathcal{S}_h(s, a_h^*(s))$,

$$\tilde{c}_h^k(s, a_h^*(s), s'(s, a_h^*(s))) > \bar{c} - \Delta_h(s, a_h^*(s), s'(s, a_h^*(s))) - 2\beta \left\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s, a, \tilde{s}'_{\max}))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}. \tag{51}$$

However, as we discussed in our Idea II and Idea III in Section 3, due to possible unsafe transitions and unsafe states in our problem, such a safety value in (51) may not be achieved by the algorithm. This is a critical difference compared with the case without instantaneous constraints or with only unsafe actions. Therefore, in the following, we first quantify the gap between the state-action pair (s, a_0') that achieves the safety value in (51) and the optimal state-action pair $(s, a_h^*(s))$. Then, we quantify the smallest gap between the *safe* state-action pair (s, a_0) and such a possibly unsafe state-action pair (s, a_0'). Specifically, for the state-action pair (s, a_0') that takes the safety value in (51), from (49), we have

$$\alpha_{s'(s,a_0')} \le 1 - \frac{\bar{c} - c_h^0 - \Delta_h(s, a_h^*(s), s'(s, a_h^*(s))) - 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}}{\bar{c} - c_h^0 - \Delta_h(s, a_h^*(s), s'(s, a_h^*(s))) + 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}}. \tag{52}$$

Since the right-hand-side of (52) increases with $\Delta_h(s, a_h^*(s), s'(s, a_h^*(s)))$, we have

$$\alpha_{s'(s,a_0')} \le 1 - \frac{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}}. \tag{53}$$

Note that (53) quantifies the gap between the state-action pair (s, a_0') that achieves the safety value in (51) and the optimal state-action pair $(s, a_h^*(s))$. Next, we quantify the smallest gap between the safe state-action pair (s, a_0) and such a possibly unsafe state-action pair (s, a_0'). According to (53), there must exists a safe action $a_{h',0}'$ for only step h', s.t.,

$$\alpha_{s'(s,a_{h',0}')} \le 1 - \frac{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}. \tag{54}$$

Then, let $\hat{f}_h(s,a) \triangleq f_h(\boldsymbol{\phi}(s,a,\cdot) - \boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0))$ denote the normalized $\mathcal{L}_2$-distance between the features of the transitions associated with the state-action pair (s,a) and the known safe feature $\boldsymbol{\phi}(s_h^0, a_h^0, s_{h+1}^0)$. According to (54) and (19), there must exists an action a_0 that induces at least one safe subsubgraph $G_h^{k,\text{safe}}(s, a_0)$, s.t.,

$$\frac{\hat{f}_h(s,a_0)}{\hat{f}_h(s,a_0')} \geq \frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*(s), s') \in \mathcal{G}_h(s)\}} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*(s), s') \in \mathcal{G}_h(s)\}} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}. \tag{55}$$

Finally, by combining (53) and (55), since the subgraph feature space is convex, we have that the left-hand-side of (28) can be upper-bounded as follows:

$$\begin{aligned}
\tilde{f}_h(s, a_0 | s_h^* = s) &= 1 - \frac{\hat{f}_h(s, a_0 | s_h^* = s)}{\hat{f}_h(s, a_0' | s_h^* = s)} \cdot \hat{f}_h(s, a_0' | s_h^* = s) \\
&\leq 1 - \frac{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}} \\
&\qquad \cdot \frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*(s), s') \in \mathcal{G}_h(s)\}} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*(s), s') \in \mathcal{G}_h(s)\}} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}.
\end{aligned}$$

□

C. Proof of Lemma 3

As we mentioned in Section 4.2, compared with Lemma 2, the main difference in Lemma 3 is that Lemma 3 quantifies the impacts from past steps, i.e., $h' \leq h$. This special new impact results in the bonus term with parameter ϵ_4 in our Idea IV in Section 3.

Notice that Lemma 3 implies that when k increases, the UCB terms l_3 decreases to be closer to 0, and thus the right-hand-side (29) get closer to 0. Then, $\tilde{f}_h(\hat{s}, a_0)$ on the left-hand-side of (29) gets closer to 0. Notice that $\tilde{f}_h(\hat{s}, a_0)$ represents the gap between the decision of the policy π^k used by LSVI-NEW and the optimal decision. In addition, on the right-hand-side of (29), l_3 characterizes the uncertainty from past steps. Thus, the above implication from Lemma 3 is consistent with the intuition that as more safety values revealed, we should be able to get closer to the optimal action.

In this section, we provide the complete proof for Lemma 3. Please see Appendix D for our discussions and proofs on how this special new impact from past steps results in a new bonus term in our Idea IV in Section 3 and how it affects the requirements for choosing the parameters ϵ_4.

Proof. According to Lemma 5 and (54), there must exists a safe action $a_{h',0}$ at step $h' \leq h$, s.t.,

$$\alpha_{s'(s, a_{h',0})} \leq 1 - \frac{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}. \tag{56}$$

Then, according to Assumption 4, there must exists a safe action a_0 at step h, s.t.,

$$\alpha_{s'(\hat{s}, a_0)} \leq 1 - \frac{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}{\delta \left(\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}\right)}. \tag{57}$$

Finally, since $\tilde{f}_h(\hat{s}, a_0) \leq \alpha_{s'(\hat{s}, a_0)}$, we have

$$\tilde{f}_h(\hat{s}, a_0) \leq 1 - \frac{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}{\delta \left(\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}\right)}.$$

□

D. Proof of Lemma 1

In this section, we provide the proof of Lemma 1. The proof replies on Lemma 2 and Lemma 3. Recall from Section 4.2 that invariant (i) shows that, if the optimal state has been found, the estimated V-value must be higher than the optimal V-value. From a high-level point of view, if the optimal safe action has also been found, invariant (i) trivially holds. If it has not been found, thanks to our new bonus terms that essentially capture the distance between the estimated safe actions and the optimal action, invariant (i) still holds. Moreover, invariant (ii) shows that, if the optimal state has not been found, the V-value of the sub-optimal state in $\tilde{\mathcal{S}}_h^k$ is still larger than the optimal V-value. This is intuitively because $\tilde{\mathcal{S}}_h^k$ only contains safe states with transitions close enough (within a small gap captured by the small constant $\bar{\alpha}_0$) to the optimal transitions, and the distance is captured by our new bonus terms.

Proof. We prove Lemma 1 by mathematical induction.

(i) Base case: when $h = H + 1$, both invariants are trivially true, since $V_h^*(s) = V_h^k(s) = 0$.

(ii) Induction step: we hypothesize that the two invariants are true when $h = h_0$. Then, we prove that they are true for $h = h_0 - 1$.

(ii-a) Step-a: note that invariant (i) trivially holds for $h = H$ since $V_h^*(s) = V_h^k(s) = r_H(s)$. Next, we prove invariant (i) for $h < H$ by considering the following two cases, based on whether the optimal action $a_h^*(s)$ has been *found* in $\tilde{\mathcal{A}}_h^k(s)$ and *chosen* or not.

(ii-a-1) Case-1: If the optimal action $a_h^*(s)$ has been found in $\tilde{\mathcal{A}}_h^k(s)$ and chosen by π^k, i.e., $a_h^k(s) = a_h^*(s)$, based on Section D.4 in (Jia et al., 2020), we have

$$V_h^k(s) = Q_h^k(s, a_h^k(s)) = Q_h^k(s, a_h^*(s)) \geq Q_h^*(s, a_h^*(s)) = V_h^*(s), \tag{58}$$

where the inequality is because of the definition of $V_h^k(s)$ in (23) and the induction hypothesis of invariant (i) at step h_0. Notice that this step is different from the analysis in the case without constraints or with only unsafe actions. Here, the optimal action $a_h^*(s)$ must already be chosen, i.e., it is not enough to simply find that the action is safe. This is because, if the optimal action $a_h^*(s)$ is simply found to be safe while not chosen by the algorithm, a future subsubgraph that is completely different from that of the optimal policy could be visited by π^k.

(ii-a-2) Case-2: If the optimal action $a_h^*(s)$ has not been *chosen* by π^k, i.e., $a_h^k(s) \neq a_h^*(s)$, we consider the following two subcases based on whether the optimal action $a_h^*(s)$ has been *found* in $\tilde{\mathcal{A}}_h^k(s)$ or not.

(ii-a-2-I) Subcase-2-I: If the optimal action $a_h^*(s)$ has been found in $\tilde{\mathcal{A}}_h^k(s)$ by π^k, i.e., $a_h^*(s) \in \tilde{\mathcal{A}}_h^k(s)$, we have

$$V_h^k(s) = \max_{a \in \tilde{\mathcal{A}}_h^k(s)} Q_h^k(s, a) = Q_h^k(s, a_h^*(s) | V_{h+1}^k) \geq Q_h^*(s, a_h^*(s) | V_{h+1}^k) \geq Q_h^*(s, a_h^*(s)) = V_h^*(s), \tag{59}$$

where the second inequality is because of the definition of $V_h^k(s)$ in (23) and the induction hypothesis of invariant (ii) at step h_0. Recal from (11) that $Q_h^*(s, a) = r_h(s, a) + \langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^*}(s, a) \rangle$, which depends on the V-value V_{h+1}^* at next step. Thus, we write such a dependency explicitly for Q_h^k and Q_h^* in (59).

(ii-a-2-II) Subcase-2-II: If the optimal action $a_h^*(s)$ has not been found in $\tilde{\mathcal{A}}_h^k(s)$ by π^k, i.e., $a_h^*(s) \notin \tilde{\mathcal{A}}_h^k(s)$, we consider the following two subsubcases, based on the reason the optimal action $a_h^*(s)$ has not been found in $\tilde{\mathcal{A}}_h^k(s)$ by π^k.

(ii-a-2-II-A) Subsubcase-2-II-A: If the optimal action $a_h^*(s)$ has not been found in $\tilde{\mathcal{A}}_h^k(s)$ by π^k because condition 1 in (9) is violated, we have

$$\max_{s' \in \mathcal{S}_h(s, a_h^*(s))} \tilde{c}_h^k(s, a_h^*(s), s') > \bar{c},$$

Note that $V_h^k(s) = \max_{a \in \tilde{\mathcal{A}}_h^k(s)} Q_h^k(s, a) \geq Q_h^k(s, a_0)$ and the bonus term $\epsilon_4 \cdot \max_{s' \in \mathcal{S}_1(s_1, a_1^k)} \|\boldsymbol{\psi}(\mathcal{U}_1^{\perp}, \boldsymbol{\phi}(s_1, a_1^k, s'))\|_{(\boldsymbol{\Lambda}_{1,1}^k)^{-1}}$ in (13) is non-negative, we have

$$\begin{aligned} V_h^k(s) \geq \min \Big\{ & r_h(s, a_0) + \Big\langle \boldsymbol{w}_h^k, \boldsymbol{\phi}_{V_{h+1}^k}(s, a_0) \Big\rangle + \epsilon_1 \cdot \|\boldsymbol{\phi}_{V_{h+1}^k}(s, a_0)\|_{(\boldsymbol{\Lambda}_{h,2}^k)^{-1}} \\ & + \epsilon_{h,2} \cdot \max_{s' \in \mathcal{S}_h(s, a_0)} \|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s, a_0, s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} + \epsilon_{h,3} \cdot \max_{(s_{h'}, a_{h'}, s') \in \mathcal{G}_h(s)} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}, a_{h'}, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}, H \Big\}. \end{aligned}$$

Then, according to Section D.4 in (Jia et al., 2020), we have

$$\begin{aligned} V_h^k(s) \geq \min\Big\{ & r_h(s,a_0) + \Big\langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^k}(s,a_0)\Big\rangle + (\epsilon_1 - 1)\cdot \|\boldsymbol{\phi}_{V_{h+1}^k}(s,a_0)\|_{(\boldsymbol{\Lambda}_{h,2}^k)^{-1}} \\ & + \epsilon_{h,2}\cdot \max_{s'\in\mathcal{S}_h(s,a_0)} \|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a_0,s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} + \epsilon_{h,3}\cdot \max_{(s_{h'},a_{h'},s')\in\mathcal{G}_h(s)} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'},a_{h'},s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}, H\Big\}. \end{aligned} \tag{60}$$

Moreover, according to Lemma 2, there must exists an action $a_0 \in \tilde{\mathcal{A}}_h^k(s)$, s.t.,

$$\begin{aligned} \tilde{f}_h(s,a_0|s_h^*=s) \leq 1 - & \frac{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta\max_{s'}\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a_h^*(s),s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta\max_{s'}\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a_h^*(s),s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}} \\ & \cdot \frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta\max_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*(s), s')\in\mathcal{G}_h(s)\}}\|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta\max_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*(s), s')\in\mathcal{G}_h(s)\}}\|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}. \end{aligned} \tag{61}$$

By combining (60) and (61), and according to Assumption 4 and invariant (ii) at the next step h_0, we have

$$\begin{aligned} V_h^k(s) \geq \min\Bigg\{ & \frac{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta\max_{s'}\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a_h^*(s),s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta\max_{s'}\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a_h^*(s),s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}} \\ & \cdot \frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta\max_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*(s), s')\in\mathcal{G}_h(s)\}}\|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta\max_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*(s), s')\in\mathcal{G}_h(s)\}}\|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}} \cdot \delta\Big[r_h(s,a_h^*(s)) \\ & + \Big\langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^*}(s,a_h^*(s))\Big\rangle + (\epsilon_1 - 1)\cdot\|\boldsymbol{\phi}_{V_{h+1}^*}(s,a_h^*(s))\|_{(\boldsymbol{\Lambda}_{h,2}^k)^{-1}} + \epsilon_{h,2} \\ & \cdot \max_{s'\in\mathcal{S}_h(s,a_h^*(s))}\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a_h^*(s),s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} + \epsilon_{h,3}\cdot \max_{(s_{h'}^*,a_{h'}^*,s')\in\mathcal{G}_h(s)}\|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}\Big], H\Bigg\}. \end{aligned}$$

Since ϵ_1 is set to be equal to $\beta+1$, we have $\epsilon_1 - 1 \geq 0$. Thus, $(\epsilon_1 - 1)\cdot\|\boldsymbol{\phi}_{V_{h+1}^*}(s,a_h^*(s))\|_{(\boldsymbol{\Lambda}_{h,2}^k)^{-1}} \geq 0$. Thus, we have

$$\begin{aligned} V_h^k(s) \geq \min\Bigg\{ & \frac{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta\max_{s'}\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a_h^*(s),s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta\max_{s'}\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a_h^*(s),s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}} \\ & \cdot \frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta\max_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*(s), s')\in\mathcal{G}_h(s)\}}\|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta\max_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*(s), s')\in\mathcal{G}_h(s)\}}\|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}} \cdot \delta\Big[r_h(s,a_h^*(s)) \\ & + \Big\langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^*}(s,a_h^*(s))\Big\rangle + \epsilon_{h,2}\cdot \max_{s'\in\mathcal{S}_h(s,a_h^*(s))}\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a_h^*(s),s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} \\ & + \epsilon_{h,3}\cdot \max_{(s_{h'}^*,a_{h'}^*,s')\in\mathcal{G}_h(s)}\|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}\Big], H\Bigg\}. \end{aligned} \tag{62}$$

Thus, to prove that $V_h^k(s) \geq V_h^*(s)$, we need to prove that

$$\begin{aligned} & \delta\Big[\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta\max_{s'}\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a_h^*(s),s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}\Big] \\ & \quad \cdot \Big[\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*(s), s')\in\mathcal{G}_h(s)\}}\|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}\Big] \\ & \quad \cdot \Big[Q_h^*(s,a_h^*(s)) + \epsilon_{h,2}\cdot \max_{s'\in\mathcal{S}_h(s,a_h^*(s))}\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a_h^*(s),s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} \\ & \qquad\qquad + \epsilon_{h,3}\cdot \max_{(s_{h'}^*,a_{h'}^*,s')\in\mathcal{G}_h(s)}\|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}\Big] \\ & \geq \Big[\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta\max_{s'}\|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s,a_h^*(s),s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}\Big] \\ & \quad \cdot \Big[\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*(s), s')\in\mathcal{G}_h(s)\}}\|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}\Big]\cdot Q_h^*(s,a_h^*(s)). \end{aligned} \tag{63}$$

By rearranging the terms in (63), we have

$$
\begin{aligned}
&\delta\Big[\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}\Big] \\
&\quad \cdot \Big[\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{\{h<h'\le H, (s_{h'}^*, a_{h'}^*(s), s') \in \mathcal{G}_h(s)\}} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}\Big] \\
&\quad \cdot \Big[\epsilon_{h,2} \cdot \max_{s' \in \mathcal{S}_h(s, a_h^*(s))} \|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} + \epsilon_{h,3} \cdot \max_{(s_{h'}^*, a_{h'}^*, s') \in \mathcal{G}_h(s)} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}\Big] \\
&\ge 4\beta\Big[(\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c)) \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} \\
&\qquad\quad + (\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c)) \max_{\{h<h'\le H, (s_{h'}^*, a_{h'}^*(s), s') \in \mathcal{G}_h(s)\}} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}\Big] \cdot Q_h^*(s, a_h^*(s)).
\end{aligned}
$$

Since $Q_h^*(s, a_h^*(s)) \le H$ for all states s and steps h, we have

$$
\begin{aligned}
&\epsilon_{h,2} \cdot \max_{s' \in \mathcal{S}_h(s, a_h^*(s))} \|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} + \epsilon_{h,3} \cdot \max_{(s_{h'}^*, a_{h'}^*, s') \in \mathcal{G}_h(s)} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}} \\
&\ge \frac{4\beta H}{\delta}\Bigg[\frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c)}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c)} \max_{s'} \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s, a_h^*(s), s'))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} \\
&\qquad\qquad\qquad + \max_{\{h<h'\le H, (s_{h'}^*, a_{h'}^*(s), s') \in \mathcal{G}_h(s)\}} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}\Bigg] \\
&\cdot \Bigg[\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - \frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c)}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c)} 2\beta \max_{s'} \left\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s, a_h^*(s), s'))\right\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} \\
&\qquad\qquad\qquad - 2\beta \max_{\{h<h'\le H, (s_{h'}^*, a_{h'}^*(s), s') \in \mathcal{G}_h(s)\}} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}\Bigg]^{-1}.
\end{aligned}
\tag{64}
$$

Note that (64) indicates that, to have $V_h^k(s) \ge V_h^*(s)$, we need

$$
\epsilon_{h,2} \ge \frac{4\beta H \frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c)}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c)}}{\delta(\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - \frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c)}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c)} \kappa)} \text{ and } \epsilon_{h,3} \ge \frac{4\beta H}{\delta(\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - \kappa)}.
$$

This is reason we set the parameters $\epsilon_{h,2}$ and $\epsilon_{h,3}$ in our Idea II and Idea III to be in the form in (14) and (15), respectively.

(ii-a-2-II-B) Subsubcase-2-II-B: If the optimal action $a_h^*(s)$ has not been found in $\tilde{\mathcal{A}}_h^k(s)$ by π^k because (although condition 1 in (9) is satisfied) condition 2 in (10) is violated, we have

$$
\mathcal{S}_{h+1}(s, a_h^*(s)) \not\subseteq \mathcal{S}_{h+1}^{k,\text{safe}}.
$$

In this subsubcase, we can leverage the knowledge from the satisfied condition 1 to prove $V_h^k(s) \ge V_h^*(s)$. The proof then could follow the similar inductions in the proof for subsubcase-2-II-A. For completeness, we provide the proof steps below. First, since the bonus term $\epsilon_4 \cdot \max_{s' \in \mathcal{S}_1(s_1, a_1^k)} \|\boldsymbol{\psi}(\mathcal{U}_1^{\perp}, \boldsymbol{\phi}(s_1, a_1^k, s'))\|_{(\boldsymbol{\Lambda}_{1,1}^k)^{-1}}$ in (13) is non-negative, according to Section D.4 in (Jia et al., 2020), we have

$$
\begin{aligned}
V_h^k(s) \ge \min\Big\{&r_h(s, a_0) + \Big\langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^k}(s, a_0)\Big\rangle + (\epsilon_1 - 1) \cdot \|\boldsymbol{\phi}_{V_{h+1}^k}(s, a_0)\|_{(\boldsymbol{\Lambda}_{h,2}^k)^{-1}} \\
&+ \epsilon_{h,2} \cdot \max_{s' \in \mathcal{S}_h(s, a_0)} \|\boldsymbol{\psi}(\mathcal{U}_h^{\perp}, \boldsymbol{\phi}(s, a_0, s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}} + \epsilon_{h,3} \cdot \max_{(s_{h'}, a_{h'}, s') \in \mathcal{G}_h(s)} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}, a_{h'}, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}, H\Big\}.
\end{aligned}
$$

Next, according to Assumption 4, invariant (ii) at next step $h+1$ and $(\epsilon_1 - 1) \cdot \|\boldsymbol{\phi}_{V^*_{h+1}}(s, a^*_h(s))\|_{(\boldsymbol{\Lambda}^k_{h,2})^{-1}} \geq 0$, we have

$$
\begin{aligned}
V^k_h(s) \geq \min \Bigg\{ & \frac{\bar{c} - c^0_h - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}^\perp_h, \boldsymbol{\phi}(s, a^*_h(s), s'))\|_{(\boldsymbol{\Lambda}^k_{h,1})^{-1}}}{\bar{c} - c^0_h - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}^\perp_h, \boldsymbol{\phi}(s, a^*_h(s), s'))\|_{(\boldsymbol{\Lambda}^k_{h,1})^{-1}}} \\
& \cdot \frac{\bar{c} - \bar{c}^0_{h'} - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{\{h<h'\leq H, (s^*_{h'}, a^*_{h'}(s), s') \in \mathcal{G}_h(s)\}} \|\boldsymbol{\psi}(\mathcal{U}^\perp_{h'}, \boldsymbol{\phi}(s^*_{h'}, a^*_{h'}(s), s'))\|_{(\boldsymbol{\Lambda}^k_{h',1})^{-1}}}{\bar{c} - \bar{c}^0_{h'} - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{\{h<h'\leq H, (s^*_{h'}, a^*_{h'}(s), s') \in \mathcal{G}_h(s)\}} \|\boldsymbol{\psi}(\mathcal{U}^\perp_{h'}, \boldsymbol{\phi}(s^*_{h'}, a^*_{h'}(s), s'))\|_{(\boldsymbol{\Lambda}^k_{h',1})^{-1}}} \cdot \delta \Big[r_h(s, a^*_h(s)) \\
& + \Big\langle w^*_h, \boldsymbol{\phi}_{V^*_{h+1}}(s, a^*_h(s)) \Big\rangle + \epsilon_{h,2} \cdot \max_{s' \in \mathcal{S}_h(s, a^*_h(s))} \|\boldsymbol{\psi}(\mathcal{U}^\perp_h, \boldsymbol{\phi}(s, a^*_h(s), s'))\|_{(\boldsymbol{\Lambda}^k_{h,1})^{-1}} \\
& + \epsilon_{h,3} \cdot \max_{(s^*_{h'}, a^*_{h'}, s') \in \mathcal{G}_h(s)} \|\boldsymbol{\psi}(\mathcal{U}^\perp_{h'}, \boldsymbol{\phi}(s^*_{h'}, a^*_{h'}, s'))\|_{(\boldsymbol{\Lambda}^k_{h',1})^{-1}} \Big], H \Bigg\}.
\end{aligned}
$$

Then, to prove $V^k_h(s) \geq V^*_h(s)$, based on (63) and since $Q^*_h(s) \leq H$, we have

$$
\begin{aligned}
& \epsilon_{h,2} \cdot \max_{s' \in \mathcal{S}_h(s, a^*_h(s))} \|\boldsymbol{\psi}(\mathcal{U}^\perp_h, \boldsymbol{\phi}(s, a^*_h(s), s'))\|_{(\boldsymbol{\Lambda}^k_{h,1})^{-1}} + \epsilon_{h,3} \cdot \max_{(s^*_{h'}, a^*_{h'}, s') \in \mathcal{G}_h(s)} \|\boldsymbol{\psi}(\mathcal{U}^\perp_{h'}, \boldsymbol{\phi}(s^*_{h'}, a^*_{h'}, s'))\|_{(\boldsymbol{\Lambda}^k_{h',1})^{-1}} \\
& \geq \frac{4\beta H}{\delta} \Bigg[\frac{\bar{c} - \bar{c}^0_{h'} - \Delta_{\boldsymbol{\phi}}(c)}{\bar{c} - c^0_h - \Delta_{\boldsymbol{\phi}}(c)} \max_{s'} \left\| \boldsymbol{\psi}(\mathcal{U}^\perp_h, \boldsymbol{\phi}(s, a^*_h(s), s')) \right\|_{(\boldsymbol{\Lambda}^k_{h,1})^{-1}} \\
& \qquad\qquad + \max_{\{h<h'\leq H, (s^*_{h'}, a^*_{h'}(s), s') \in \mathcal{G}_h(s)\}} \|\boldsymbol{\psi}(\mathcal{U}^\perp_{h'}, \boldsymbol{\phi}(s^*_{h'}, a^*_{h'}(s), s'))\|_{(\boldsymbol{\Lambda}^k_{h',1})^{-1}} \Bigg] \\
& \cdot \Bigg[\bar{c} - \bar{c}^0_{h'} - \Delta_{\boldsymbol{\phi}}(c) - \frac{\bar{c} - \bar{c}^0_{h'} - \Delta_{\boldsymbol{\phi}}(c)}{\bar{c} - c^0_h - \Delta_{\boldsymbol{\phi}}(c)} 2\beta \max_{s'} \left\| \boldsymbol{\psi}(\mathcal{U}^\perp_h, \boldsymbol{\phi}(s, a^*_h(s), s')) \right\|_{(\boldsymbol{\Lambda}^k_{h,1})^{-1}} \\
& \qquad\qquad - 2\beta \max_{\{h<h'\leq H, (s^*_{h'}, a^*_{h'}(s), s') \in \mathcal{G}_h(s)\}} \|\boldsymbol{\psi}(\mathcal{U}^\perp_{h'}, \boldsymbol{\phi}(s^*_{h'}, a^*_{h'}(s), s'))\|_{(\boldsymbol{\Lambda}^k_{h',1})^{-1}} \Bigg]^{-1},
\end{aligned}
$$

which provides the same requirements on the parameters $\epsilon_{h,2}$ and $\epsilon_{h,3}$.

(ii-b) Step-b: differently from invariant (i) that trivially holds for $h = H$, we need to carefully handle the correctness of invariant (ii) at step $h = H$. Next, we prove invariant (ii) for all steps $h \leq H$ as follows.

First, since the bonus terms $\epsilon_{h,2} \cdot \max_{s' \in \mathcal{S}_h(s,a)} \|\boldsymbol{\psi}(\mathcal{U}^\perp_h, \boldsymbol{\phi}(s, a, s'))\|_{(\boldsymbol{\Lambda}^k_{h,1})^{-1}}$ and $\epsilon_{h,3} \cdot \max_{(s_{h'}, a_{h'}, s') \in \mathcal{G}_h(s)} \|\boldsymbol{\psi}(\mathcal{U}^\perp_{h'}, \boldsymbol{\phi}(s_{h'}, a_{h'}, s'))\|_{(\boldsymbol{\Lambda}^k_{h',1})^{-1}}$ in (13) are non-negative, to prove $V^k_h(\hat{s}) \geq V^*_h(s)$, we need to prove that

$$
\begin{aligned}
& r_h(\hat{s}, \hat{a}) + \langle \boldsymbol{w}^k_h, \boldsymbol{\phi}_{V^k_{h+1}}(\hat{s}, \hat{a}) \rangle + \epsilon_1 \cdot \|\boldsymbol{\phi}_{V^k_{h+1}}(\hat{s}, \hat{a})\|_{(\boldsymbol{\Lambda}^k_{h,2})^{-1}} + \epsilon_4 \cdot \max_{s' \in \mathcal{S}_1(s_1, a^k_1)} \|\boldsymbol{\psi}(\mathcal{U}^\perp_1, \boldsymbol{\phi}(s_1, a^k_1, s'))\|_{(\boldsymbol{\Lambda}^k_{1,1})^{-1}} \\
& \geq r_h(s, a^*_h(s)) + \Big\langle w^*_h, \boldsymbol{\phi}_{V^*_{h+1}}(s, a^*_h(s)) \Big\rangle,
\end{aligned} \tag{65}
$$

for some $\hat{a} \in \tilde{\mathcal{A}}^k_h(\hat{s})$. To prove (65), we prove

$$
\begin{aligned}
& \Big[r_h(s, a^*_h(s)) + \Big\langle w^*_h, \boldsymbol{\phi}_{V^*_{h+1}}(s, a^*_h(s)) \Big\rangle \Big] - \Big[r_h(\hat{s}, \hat{a}) + \langle \boldsymbol{w}^k_h, \boldsymbol{\phi}_{V^k_{h+1}}(\hat{s}, \hat{a}) \rangle \Big] \\
& \leq \epsilon_1 \cdot \|\boldsymbol{\phi}_{V^k_{h+1}}(\hat{s}, \hat{a})\|_{(\boldsymbol{\Lambda}^k_{h,2})^{-1}} + \epsilon_4 \cdot \max_{s' \in \mathcal{S}_1(s_1, a^k_1)} \|\boldsymbol{\psi}(\mathcal{U}^\perp_1, \boldsymbol{\phi}(s_1, a^k_1, s'))\|_{(\boldsymbol{\Lambda}^k_{1,1})^{-1}}.
\end{aligned} \tag{66}
$$

By adding and subtracting $r_h(\hat{s}, \hat{a}) + \langle w^*_h, \boldsymbol{\phi}_{V^*_{h+1}}(\hat{s}, \hat{a}) \rangle$, we decompose the left-hand-side of (66) into two parts that are easier for analysis in the following special way,

$$
\begin{aligned}
& \Big[r_h(s, a^*_h(s)) + \Big\langle w^*_h, \boldsymbol{\phi}_{V^*_{h+1}}(s, a^*_h(s)) \Big\rangle \Big] - \Big[r_h(\hat{s}, \hat{a}) + \langle \boldsymbol{w}^k_h, \boldsymbol{\phi}_{V^k_{h+1}}(\hat{s}, \hat{a}) \rangle \Big] \\
& = \Big[r_h(s, a^*_h(s)) + \Big\langle w^*_h, \boldsymbol{\phi}_{V^*_{h+1}}(s, a^*_h(s)) \Big\rangle \Big] - \Big[r_h(\hat{s}, \hat{a}) + \langle w^*_h, \boldsymbol{\phi}_{V^*_{h+1}}(\hat{s}, \hat{a}) \rangle \Big] \\
& \qquad + \Big[r_h(\hat{s}, \hat{a}) + \langle w^*_h, \boldsymbol{\phi}_{V^*_{h+1}}(\hat{s}, \hat{a}) \rangle \Big] - \Big[r_h(\hat{s}, \hat{a}) + \langle \boldsymbol{w}^k_h, \boldsymbol{\phi}_{V^k_{h+1}}(\hat{s}, \hat{a}) \rangle \Big].
\end{aligned} \tag{67}
$$

Notice that by decomposing in this way, the value in the first two brackets $[\cdot]$ on the right-hand-side of (67) characterizes how the policy executed by our LSVI-NEW algorithm learns about and searches towards the optimal safe subgraph. The value in the last two brackets $[\cdot]$ on the right-hand-side of (67) characterizes how the policy executed by our LSVI-NEW algorithm learns and estimates the optimal Q-value parameter w_h^*. Next, according to invariant (ii) at next step h_0, the value in the last two brackets $[\cdot]$ on the right-hand-side of (67) can be upper-bounded as follows,

$$\begin{aligned}&\left[r_h(\hat{s},\hat{a}) + \langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^*}(\hat{s},\hat{a})\rangle\right] - \left[r_h(\hat{s},\hat{a}) + \langle \boldsymbol{w}_h^k, \boldsymbol{\phi}_{V_{h+1}^k}(\hat{s},\hat{a})\rangle\right] \\ &\quad \le \left[r_h(\hat{s},\hat{a}) + \langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^k}(\hat{s},\hat{a})\rangle\right] - \left[r_h(\hat{s},\hat{a}) + \langle \boldsymbol{w}_h^k, \boldsymbol{\phi}_{V_{h+1}^k}(\hat{s},\hat{a})\rangle\right] \le \epsilon_1 \cdot \|\boldsymbol{\phi}_{V_{h+1}^k}(\hat{s},\hat{a})\|_{(\mathbf{\Lambda}_{h,2}^k)^{-1}}.\end{aligned}$$

Then, to prove (66), we need to prove

$$\begin{aligned}&\left[r_h(s, a_h^*(s)) + \left\langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^*}(s, a_h^*(s))\right\rangle\right] - \left[r_h(\hat{s},\hat{a}) + \langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^*}(\hat{s},\hat{a})\rangle\right] \\ &\quad \le \epsilon_4 \cdot \max_{s' \in \mathcal{S}_1(s_1, a_1^k)} \|\boldsymbol{\psi}(\mathcal{U}_1^{\perp}, \boldsymbol{\phi}(s_1, a_1^k, s'))\|_{(\mathbf{\Lambda}_{1,1}^k)^{-1}}.\end{aligned} \tag{68}$$

Therefore, below we focus on bounding the value in the first two brackets on the right-hand-side of (67). According to the definition of $V_h^k(s)$ and Lemma 3, there must exist an action $\hat{a} \in \tilde{\mathcal{A}}_h^k(\hat{s})$ and $1 \le h' \le h$, s.t.,

$$\tilde{f}_h(\hat{s},\hat{a}) \le 1 - \frac{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max\limits_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\mathbf{\Lambda}_{h',1}^k)^{-1}}}{\delta(\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max\limits_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\mathbf{\Lambda}_{h',1}^k)^{-1}})}.$$

Thus, we have

$$\begin{aligned}&r_h(\hat{s},\hat{a}) + \left\langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^*}(\hat{s},\hat{a})\right\rangle \\ &\ge \frac{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max\limits_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\mathbf{\Lambda}_{h',1}^k)^{-1}}}{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max\limits_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\mathbf{\Lambda}_{h',1}^k)^{-1}}} \left[r_h(s, a_h^*(s)) + \left\langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^*}(s, a_h^*(s))\right\rangle\right].\end{aligned} \tag{69}$$

Notice that (69) indicates that the left-hand-side of (68) can be upper-bounded as follows,

$$\begin{aligned}&\left[r_h(s, a_h^*(s)) + \left\langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^*}(s, a_h^*(s))\right\rangle\right] - \left[r_h(\hat{s},\hat{a}) + \langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^*}(\hat{s},\hat{a})\rangle\right] \\ &\le \frac{4\beta \max\limits_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\mathbf{\Lambda}_{h',1}^k)^{-1}}}{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max\limits_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\mathbf{\Lambda}_{h',1}^k)^{-1}}} \left[r_h(s, a_h^*(s)) + \left\langle w_h^*, \boldsymbol{\phi}_{V_{h+1}^*}(s, a_h^*(s))\right\rangle\right] \\ &\le \frac{4\beta H \max\limits_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\mathbf{\Lambda}_{h',1}^k)^{-1}}}{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max\limits_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\mathbf{\Lambda}_{h',1}^k)^{-1}}},\end{aligned} \tag{70}$$

where the last inequality is because $V_h^*(s) \le H$ for all states s and steps h. (70) indicates that to prove (68), we need

$$\begin{aligned}&(\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\mathbf{\Lambda}_{h',1}^k)^{-1}}) \cdot \epsilon_4 \cdot \max_{s' \in \mathcal{S}_1(s_1, a_1^k)} \|\boldsymbol{\psi}(\mathcal{U}_1^{\perp}, \boldsymbol{\phi}(s_1, a_1^k, s'))\|_{(\mathbf{\Lambda}_{1,1}^k)^{-1}} \\ &\ge 4\beta H \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp}, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\mathbf{\Lambda}_{h',1}^k)^{-1}}.\end{aligned} \tag{71}$$

Note that (71) shows that, to prove $V_h^k(\hat{s}) \ge V_h^*(s)$, we need

$$\epsilon_4 \ge \frac{4\beta H}{\bar{c} - c_1^0 - \Delta_{\boldsymbol{\phi}}(c)}.$$

This is the reason we set the parameter ϵ_4 in our Idea IV to be in the form in (16).

□

E. Proof of Theorem 2

As we mentioned in Section 4.2, because of the new challenges from the instantaneous hard constraint (1) and our novel ideas in the algorithm design, there are several new difficulties in the regret analysis, which is shown in this section. The key ones are: (I) Differently from the unconstrained setting or the setting with only unsafe actions, in our case, the states that could be visited with non-zero probability by different policies could be completely different at each step h. Hence, the commonly-used invariant on V-values, i.e., $V_h^k(s) \geq V_h^*(s)$ for all h and s, that relies on the ergodicity property no longer holds in our case. This difficulty is resolved by Lemma 1. (II) How to quantify the impacts when looking ahead and peeking backward. This difficulty is resolved by Lemma 2 and Lemma 3.

Proof. First, for the convenience of the reader, we restate our new construction for the V-values functions of different policies. We let $\mathcal{S}_h^*$ denote the state set at step h in the optimal safe subgraph. Let $\mathcal{S}_h^k$ denote the state set at step h in the subgraph followed by policy π^k of LSVI-NEW in episode k. Moreover, we let $\tilde{f}_h(s,a) \triangleq f_h(\boldsymbol{\phi}(s,a,\cdot) - \boldsymbol{\phi}(s_h^*, a_h^*, \cdot))$ denote the gap between the transitions associated with the state-action pair (s,a) and the optimal transitions. Let $\tilde{\mathcal{A}}_h^k(s) \triangleq \{a \in \mathcal{A}_h^{k,\text{safe}}(s) : \tilde{f}_h(s,a) \leq \bar{\alpha}_0\} \cup \{a_h^k(s)\}$ denote the union of the safe actions with transitions close to the optimal transitions and the action chosen by π^k for a safe state s at step h, where

$$\bar{\alpha}_0 = \max\left\{1 - \frac{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}}{\bar{c} - c_h^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max_{s'} \|\boldsymbol{\psi}(\mathcal{U}_h^\perp, \boldsymbol{\phi}(s, a_h^*(s), s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}} \right.$$
$$\cdot \frac{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max\limits_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*), s'\}} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}{\bar{c} - \bar{c}_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max\limits_{\{h<h'\leq H, (s_{h'}^*, a_{h'}^*), s'\}} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}},$$
$$\left. 1 - \frac{\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) - 2\beta \max\limits_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}}{\delta(\bar{c} - c_{h'}^0 - \Delta_{\boldsymbol{\phi}}(c) + 2\beta \max\limits_{s'} \|\boldsymbol{\psi}(\mathcal{U}_{h'}^\perp, \boldsymbol{\phi}(s_{h'}^*, a_{h'}^*, s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}})} \right\}$$

is a small value that decreases to be closer to 0 when the number of learning episodes k increases. Let $\tilde{\mathcal{S}}_h^k \triangleq \{s \in \mathcal{S}_h^{k,\text{safe}} : \exists a \in \mathcal{A}_h^{k,\text{safe}}(s), \text{ s.t., } \tilde{f}_h(s,a) \leq \bar{\alpha}_0\} \cup \mathcal{S}_h^k$ denote the union of the safe states with transitions close to the optimal transitions and the state set at step h in the subgraph followed by policy π^k of LSVI-NEW in episode k. Next, we define the V-value functions of the optimal policy, estimated policy and policy π^k to be

$$V_h^*(s) \triangleq Q_h^*(s, a_h^*(s)), \forall s \in \mathcal{S}_h^*, \tag{72}$$
$$V_h^k(s) \triangleq \max_{a \in \tilde{\mathcal{A}}_h^k(s)} Q_h^k(s,a), \forall s \in \tilde{\mathcal{S}}_h^k, \tag{73}$$
$$V_h^{\pi^k}(s) \triangleq Q_h^{\pi^k}(s, a_h^k(s)), \forall s \in \mathcal{S}_h^k, \tag{74}$$

respectively. Then, the regret $R^{\text{LSVI-NEW}}$ can be decomposed into two parts as follows:

$$R^{\text{LSVI-NEW}} = \sum_{k=1}^K \left\{V_1^*(s_1) - V_1^{\pi^k}(s_1)\right\} = \sum_{k=1}^K \left\{\left[V_1^*(s_1) - V_1^k(s_1)\right] + \left[V_1^k(s_1) - V_1^{\pi^k}(s_1)\right]\right\}. \tag{75}$$

To upper-bound the regret, we prove that, with high probability, (i) the value in the first bracket on the right-hand-side of (25) is non-positive; (ii) the value in the second bracket on the right-hand-side of (75) can be upper-bounded. Note that, according to Lemma 1, we have the value in the first bracket on the right-hand-side of (75) must be non-positive, i.e., $V_1^*(s_1) - V_1^k(s_1) \leq 0$ for all episodes k. The value in the second bracket on the right-hand-side of (75) can be upper-bounded by slightly modifying existing techniques for the linear mixture MDP. Specifically, according to the Azuma-Hoeffding inequality, we have

$$\sum_{k=1}^{K}\left\{V_1^k(s_1)-V_1^{\pi^k}(s_1)\right\}\leq\sum_{k=1}^{K}\sum_{h=1}^{H}\Big\{\epsilon_1\cdot\|\boldsymbol{\phi}_{V_{h+1}^k}(s,a)\|_{(\boldsymbol{\Lambda}_{h,2}^k)^{-1}}+\epsilon_{h,2}\cdot\max_{s'\in\mathcal{S}_h(s,a)}\|\boldsymbol{\psi}(\mathcal{U}_h^{\perp},\boldsymbol{\phi}(s,a,s'))\|_{(\boldsymbol{\Lambda}_{h,1}^k)^{-1}}$$

$$+\epsilon_{h,3}\cdot\max_{(s_{h'},a_{h'},s')\in\mathcal{G}_h(s)}\|\boldsymbol{\psi}(\mathcal{U}_{h'}^{\perp},\boldsymbol{\phi}(s_{h'},a_{h'},s'))\|_{(\boldsymbol{\Lambda}_{h',1}^k)^{-1}}+\epsilon_4\cdot\max_{s'\in\mathcal{S}_1(s_1,a_1^k)}\|\boldsymbol{\psi}(\mathcal{U}_1^{\perp},\boldsymbol{\phi}(s_1,a_1^k,s'))\|_{(\boldsymbol{\Lambda}_{1,1}^k)^{-1}}\Big\}$$

$$+2H\sqrt{HK\log\left(\frac{2dHK}{p}\right)}+HK'$$

$$\leq\sum_{k=1}^{K}\sum_{h=1}^{H}\left\{\beta+1+\frac{\frac{4\beta H}{\tilde{\delta}}\frac{\bar{c}-\bar{c}_{h'}^0-\Delta_{\boldsymbol{\phi}}(c)}{\bar{c}-c_h^0-\Delta_{\boldsymbol{\phi}}(c)}}{\bar{c}-\bar{c}_{h'}^0-\Delta_{\boldsymbol{\phi}}(c)-\frac{\bar{c}-\bar{c}_{h'}^0-\Delta_{\boldsymbol{\phi}}(c)}{\bar{c}-c_h^0-\Delta_{\boldsymbol{\phi}}(c)}\kappa}+\frac{4\beta H/\tilde{\delta}}{\bar{c}-\bar{c}_{h'}^0-\Delta_{\boldsymbol{\phi}}(c)-\kappa}+\frac{4\beta H}{\bar{c}-c_1^0-\Delta_{\boldsymbol{\phi}}(c)}\right\}$$

$$\cdot\sqrt{2dHK\log\left(1+HK\right)}+2H\sqrt{HK\log\left(\frac{2dHK}{p}\right)}+HK'+\frac{D}{\lambda_0}\left(\frac{K}{K'}-1\right),$$

where the last inequality is because of Lemma D.2 in (Jin et al., 2020) and Lemma 1 in (Amani et al., 2019).

□

F. Proof of Theorem 3

In this section, we provide the proof for Theorem 3. The proof is based on the lower bound in the unconstrained horizon-free linear mixture MDP setting (Zhou & Gu, 2022) and the lower bound in the constrained bandit setting (Pacchiano et al., 2021). Note that these existing lower bounds do not show the dependency on the episode length H and the safety parameter $\Delta_{\boldsymbol{\phi}}(c)$ that are captured in our lower bound.

Proof. Notice that in Theorem 3, we assume $K \geq 32\underline{R}$. Under this assumption, Lemma 25 in Zhou et al. (2021a) indicates that in the linear bandit problems that are parameterized by the vector $\mu^* = \left\{-\frac{\sqrt{\delta/K}}{4\sqrt{2}}, \frac{\sqrt{\delta/K}}{4\sqrt{2}}\right\}^d$ and with the action space $\mathcal{A} = \{-1, 1\}^d$ and Bernoulli distributed reward $r \sim \mathcal{B}(\delta + \langle \mu^*, a\rangle)$, where $0 < \delta \leq \frac{1}{3}$, the regret of any algorithm is lower-bounded by $\frac{dH\sqrt{K}}{8\sqrt{2}}$. Next, consider an instance with three states $\{s_1, s_2, s_3\}$, one action a, and the reward $r_h(s_1, a) = r_h(s_2, a) = 0$ and $r_h(s_3, a) = 1$ for each h. Then, by using the same transition probability in Section C.3 of Zhou & Gu (2022), we have that the regret of any algorithm for linear mixture MDPs with H steps in each episode is lower-bounded by $\frac{dH\sqrt{K}}{16\sqrt{2}}$. Since the linear mixture MDP with instantaneous hard constraints subsumes (when the cost $c_h(s, a, s') = 0$ for all state-action-state triplets) the unconstrained case, $\frac{dH\sqrt{K}}{16\sqrt{2}}$ is also a lower bound of the regret in our case.

Further, to quantify the impact of the safety term $\bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c)$ on the lower bound, in the following, we focus on showing that, when the instantaneous hard constraint with threshold $\bar{c}$ is considered, the regret is at least $\frac{H}{24(\bar{c}-\bar{c}_1^0-\Delta_{\boldsymbol{\phi}}(c))^2}$. We prove this by contradiction. Assume there exists a safe algorithm that can achieve a regret $R_0 < \frac{H}{24(\bar{c}-\bar{c}_1^0-\Delta_{\boldsymbol{\phi}}(c))^2}$ for any instance of the problem that we consider. Let us consider the following transition probability function: At step $h = 1$, the transition probability is equal to $\mathbb{P}_1(s_2(i)|s_1, a(i)) = 1$ for all i, and $\mathbb{P}_1(s_2(i)|s_1, a(j)) = 0$ for all $i \neq j$; at step $h > 1$, the transition probability is equal to $\mathbb{P}_h(s_{h+1}(i)|s_h(i), a(j)) = 1$ for all i and j, and $\mathbb{P}_h(s_{h+1}(j)|s_h(i), a(l)) = 0$ for all $i \neq j$ and all l, where i, j and l are the indices of the states and actions.

Now, let us consider an instance where the safety value function is as follows: at step $h = 1$, the safety value is equal to $c_1(s_1, a(1), s') = \bar{c}_1^0$, $c_1(s_1, a(2), s') = 2\bar{c} - \bar{c}_1^0$, $c_1(s_1, a(3), s') = \bar{c}_1^0$, $c_1(s_1, a(4), s') = 2\bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c)$ and $c_1(s_1, a(i), s') = 2\bar{c} - \bar{c}_1^0$ for all $i > 4$. Notice that $a(1)$ and $a(3)$ are safe actions, while $a(2)$, $a(4)$ and other actions are unsafe for state s_1 at step $h = 1$. Moreover, at step $h > 1$, for all i, the safety value is equal to $c_h(s_h(1), a(i), s') = \bar{c}_1^0$, $c_1(s_h(2), a(i), s') = 2\bar{c} - \bar{c}_1^0$, $c_1(s_h(3), a(i), s') = \bar{c}_1^0$, $c_1(s_h(4), a(i), s') = 2\bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c)$ and $c_1(s_h(j), a(i), s') = 2\bar{c} - \bar{c}_1^0$ for all $j > 4$. Notice that $s_h(1)$ and $s_h(3)$ are safe states, while $s_h(2)$, $s_h(4)$ and other states are unsafe at each step $h > 1$. The reward value function is as follows: at step $h = 1$, the reward is equal to $r_1(s_1, a(1)) = \frac{1}{8}$, $r_1(s_1, a(2)) = 1$,

$r_1(s_1, a(3)) = 0$ and $r_1(s_1, a(i)) = \frac{1}{2}$ for all $i > 3$; at step $h > 1$, for all i, the reward is equal to $r_h(s_h(1), a(i)) = \frac{1}{8}$, $r_1(s_h(2), a(i)) = 1$, $r_1(s_h(3), a(i)) = 0$ and $r_1(s_h(j), a(i)) = \frac{1}{2}$ for all $j > 3$. Since for any algorithm that chooses action $a(1)$ at step $h = 1$ less than half of the total episodes with probability p_1, the regret is at least $\frac{p_1 HK}{2}$. Moreover, since the regret of assumed algorithm is $R_0 < \frac{H}{24(\bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c))^2}$, we have that, for this algorithm,

$$p_1 \leq \frac{1}{12K(\bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c))^2}.$$

Next, let us consider another instance where the safety value function is as follows: at step $h = 1$, the safety value is equal to $c_1(s_1, a(1), s') = \bar{c}_1^0$, $c_1(s_1, a(2), s') = 2\bar{c} - \bar{c}_1^0$, $c_1(s_1, a(3), s') = \bar{c}_1^0$, $c_1(s_1, a(4), s') = \bar{c}_1^0 + \Delta_{\boldsymbol{\phi}}(c)$ and $c_1(s_1, a(i), s') = 2\bar{c} - \bar{c}_1^0$ for all $i > 4$. Notice that $a(1)$, $a(3)$ and $a(4)$ are safe actions, while $a(2)$ and other actions are unsafe for state s_1 at step $h = 1$. Moreover, at step $h > 1$, for all i, the safety value is equal to $c_h(s_h(1), a(i), s') = \bar{c}_1^0$, $c_1(s_h(2), a(i), s') = 2\bar{c} - \bar{c}_1^0$, $c_1(s_h(3), a(i), s') = \bar{c}_1^0$, $c_1(s_h(4), a(i), s') = \bar{c}_1^0 + \Delta_{\boldsymbol{\phi}}(c)$ and $c_1(s_h(j), a(i), s') = 2\bar{c} - \bar{c}_1^0$ for all $j > 4$. Notice that $s_h(1)$, $s_h(3)$ and $s_h(4)$ are safe states, while $s_h(2)$ and other states are unsafe at each step $h > 1$. The reward value function is as follows: at step $h = 1$, the reward is equal to $r_1(s_1, a(1)) = \frac{1}{8}$, $r_1(s_1, a(2)) = 1$, $r_1(s_1, a(3)) = 0$ and $r_1(s_1, a(i)) = \frac{1}{2}$ for all $i > 3$; at step $h > 1$, the reward is equal to $r_h(s_h(1), a(i)) = \frac{1}{8}$, $r_1(s_h(2), a(i)) = 1$, $r_1(s_h(3), a(i)) = 0$ and $r_1(s_h(j), a(i)) = \frac{1}{2}$ for all $j > 3$. Since for any algorithm that chooses action $a(1)$ at step $h = 1$ more than half of the total episodes with probability p_2, the regret is at least $\frac{3p_2 HK}{16}$. Moreover, since the regret of the assumed algorithm is $R_0 < \frac{H}{24(\bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c))^2}$, we have, for this algorithm,

$$p_2 \leq \frac{2}{9K(\bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c))^2}.$$

Notice that the main difference between this two instances is change of the safety of action $a(4)$ for state s_1 at step $h = 1$. Specifically, in instance 1, action $a(4)$ is unsafe, while in instance 2 it becomes safe and incurs the largest reward. Thus, we can quantify the total variation distance between the statistical distributions between these two instances, which can further be upper-bounded by the Kullback–Leibler (KL) divergence. More specifically, according to Lemma 1 in Kaufmann et al. (2016) and Lemma 15.1 in Lattimore & Szepesvári (2020), we have that this KL divergence is at least $q(4) \cdot D_{\mathrm{KL}} \left(\mathcal{N}(2\bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c), \boldsymbol{I}) \| \mathcal{N}(\bar{c}_1^0 + \Delta_{\boldsymbol{\phi}}(c), \boldsymbol{I}) \right) = 2q(4)(\bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c))^2 \geq \frac{1}{2}$, where $q(4)$ is the expected number of times of choosing action $a(4)$ at step $h = 1$ in instance 1. Thus, we have

$$q(4) \geq \frac{1}{4(\bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c))^2}$$

For the algorithm choosing action $a(4)$ for at least $q(4)$ times in average for instance 1, the regret is at least $q(4) \cdot \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} q(4)$. This contradicts with our assumption that the regret of this algorithm is $R_0 < \frac{H}{24(\bar{c} - \bar{c}_1^0 - \Delta_{\boldsymbol{\phi}}(c))^2}$.

□