Long-Term Safe Reinforcement Learning with Binary Feedback

Akifumi Wachi[1], **Wataru Hashimoto**[2], **Kazumune Hashimoto**[2]

[1] LINE Corporation
[2] Osaka University
akifumi.wachi@linecorp.com, hashimoto@is.eei.eng.osaka-u.ac.jp, hashimoto@eei.eng.osaka-u.ac.jp

Abstract

Safety is an indispensable requirement for applying reinforcement learning (RL) to real problems. Although there has been a surge of safe RL algorithms proposed in recent years, most existing work typically 1) relies on receiving numeric safety feedback; 2) does not guarantee safety during the learning process; 3) limits the problem to a priori known, deterministic transition dynamics; and/or 4) assume the existence of a known safe policy for any states. Addressing the issues mentioned above, we thus propose Long-term Binary-feedback Safe RL (LoBiSaRL), a safe RL algorithm for constrained Markov decision processes (CMDPs) with binary safety feedback and an unknown, stochastic state transition function. LoBiSaRL optimizes a policy to maximize rewards while guaranteeing a long-term safety that an agent executes only safe state-action pairs throughout each episode with high probability. Specifically, LoBiSaRL models the binary safety function via a generalized linear model (GLM) and conservatively takes only a safe action at every time step while inferring its effect on future safety under proper assumptions. Our theoretical results show that LoBiSaRL guarantees the long-term safety constraint, with high probability. Finally, our empirical results demonstrate that our algorithm is safer than existing methods without significantly compromising performance in terms of reward.

1 Introduction

Safe reinforcement learning (RL) is a promising paradigm for applying RL algorithms to real-world applications (Garcıa and Fernández 2015). Safe RL is beneficial in safety-critical decision-making problems, such as autonomous driving, healthcare, and robotics, where safety requirements must be incorporated to prevent RL policies from posing risks to humans or objects (Dulac-Arnold et al. 2021). As a result, safe RL has received significant attention in recent years as a crucial issue of RL during both the learning and execution phases (Amodei et al. 2016).

Safe RL is typically formulated as *constrained* policy optimization problems where the expected cumulative reward is maximized while guaranteeing or encouraging the satisfaction of safety constraints, which are modeled as constrained Markov decision processes (CMDPs, Altman (1999)). While there are various types of constraint representations, most of the existing studies formulated constraints using either expected cumulative safety-cost (Altman 1999) or conditional value at risk (CVaR, Rockafellar, Uryasev et al. (2000)); thus, satisfying safety constraints almost surely or with high probability received less attention to date. Imagine highly safety-critical applications (e.g., autonomous driving, healthcare, robotics) where even a single constraint violation may result in catastrophic failure. In such cases, RL agents need to ensure safety at every time step at least with high probability; thus, constraint satisfaction "on average" does not fit the purpose due to a large number of unsafe actions during the learning process (Stooke, Achiam, and Abbeel 2020).

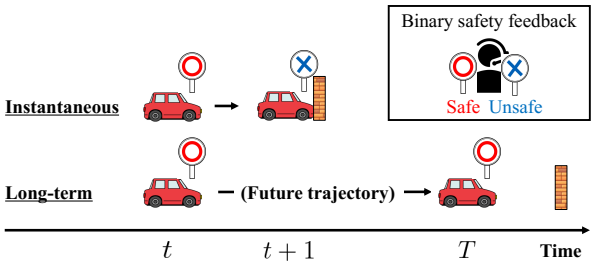


Figure 1: Even if safety is guaranteed at time t based on the instantaneous evaluation, safe behavior may not exist a few steps ahead. This paper requires an agent to guarantee long-term safety (i.e., constraint satisfaction from the time the current time step t to the terminal time step T) in CMDPs with stochastic state transition and binary safety feedback.

Several previous work on safe RL aimed to guarantee safety at every time step with high probability, even during the learning process. Unfortunately, however, existing work has room for improvement. First, most of the previous work (Wachi and Sui 2020; Amani, Thrampoulidis, and Yang 2021; Roderick, Nagarajan, and Kolter 2021) assumes numeric safety feedback. In many cases, however, the safety feedback can only take binary values indicating whether a state-action pair is safe or unsafe, which is particularly true when feedback comes from humans. As existing studies on safe RL with binary safety feedback, Wachi, Wei, and Sui (2021) modeled the safety function via a generalized lin-

	State transition		Safety	Additional assumption(s)
	Known	D/S		
Wachi and Sui (2020)	Yes	D	GP	-
Amani, Thrampoulidis, and Yang (2021)	Linear	S	Linear	Known safe policy
Wachi, Wei, and Sui (2021)	Yes	D	GLM	-
Bennett, Misra, and Kallus (2023)	No	S	GLM	Known safe policy
LoBiSaRL (Ours)	No	S	GLM	Lipschitz continuity & conservative policy

Table 1: Comparison among existing work regarding their assumptions on a state transition, safety function, and others. In the above table, D means deterministic state transition, and S means stochastic state transition.

ear model (GLM) while they assume known and deterministic state transition function. Thus, this previous work cannot deal with general RL problems with unknown stochastic state transition functions. Also, Bennett, Misra, and Kallus (2023) addressed safe RL problems with binary safety feedback and unknown stochastic state transition under the assumption that a known safe action always exists for any state. This assumption is not valid in many safety-critical applications. For example, even an F1 driver cannot take a safe action if a vehicle traveling at 100 km/h is 1 meter ahead of a brick wall; thus, to avoid such situations, we need to consider "long-term" future safety under more reasonable assumptions, as shown in Figure 1.

Contributions. We propose an algorithm called Long-term Binary-feedback Safe RL, LoBiSaRL. This algorithm enables us to solve safe RL problems with binary feedback and unknown, stochastic state transition while guaranteeing the satisfaction of long-term safety constraints. LoBiSaRL guarantees safety by modeling the binary safety function via a GLM and then pessimistically estimating the future safety function values. Our theoretical analysis shows that future safety can be pessimistically characterized by 1) inevitable randomness due to the stochastic state transition and 2) divergence between the current policy and a reference policy to stabilize the state. Based on this theoretical result, we optimize the policy to maximize the expected cumulative reward while guaranteeing long-term safety. Finally, we empirically demonstrate the effectiveness of the LoBiSaRL compared with several baselines.

2 Related Work

Safe RL. In typical safe RL problems, an agent must maximize the expected cumulative reward while ensuring that the expected cumulative cost is less than a threshold. There have been a number of algorithms for solving this type of safe RL problem, as represented by constrained policy optimization (Achiam et al. 2017), reward constrained policy optimization (Tessler, Mankowitz, and Mannor 2018), Lagrangian-based actor-critic (Chow et al. 2017), primal-dual policy optimization (Yang and Wang 2020). In the previous papers mentioned above, however, a safety constraint is defined using the (expected) cumulative value and the constraint satisfaction is *not* guaranteed during the learning process (Stooke, Achiam, and Abbeel 2020). Hence, most of the existing studies deal with less strict safety constraints than our study that requires the agent to satisfy safety constraints *at every time step*. There has been research aimed at guaranteeing safety at every time step, even during the learning process. For example, Turchetta, Berkenkamp, and Krause (2016) proposed notable algorithms that satisfy the safety constraint with high probability by inferring the safety function via a Gaussian process (GP). Also, Wachi, Wei, and Sui (2021) proposed its extended algorithm that models the safety function via a GLM, which can also deal with safe RL problems with binary feedback. Though they succeeded in guaranteeing safety with high probability, their theoretical results are based on the assumptions of the known and deterministic state transition function. It is essentially difficult to extend these algorithms to our problem settings with unknown and stochastic state transitions. As existing work on safe RL with unknown stochastic transition, Amani, Thrampoulidis, and Yang (2021) proposed an algorithm for linear MDPs with safety constraints while Bennett, Misra, and Kallus (2023) proposed an algorithm for safe RL problems with binary safety feedback and stochastic transitions. Although such work proposed excellent algorithms for challenging problems, the existence of a known safe policy is assumed for any state, which does not hold in many real-world applications as discussed in Section 1 (i.e., high-speed vehicle example). Table 1 summarizes the problem settings considered in existing work and this paper.

Long-term safety. In the control community, long-term safety has been well-studied under the name of control barrier function (CBF, Ames et al. (2019)). For any state s, a CBF is a continuously differentiable function $h(s)$ that defines a safe set $\{s : h(s) \geq 0\}$, i.e., an invariant set where any trajectory starting inside the set remains within the set. The CBF is to maintain safety during the learning process, which is particularly useful for keeping a manipulator within a given safe space or ensuring that a robot avoids obstacles. This advantage is beneficial for RL settings, and Cheng et al. (2019) proposed a safe RL algorithm to guarantee long-term safety via CBFs. Unfortunately, however, humans need to manually define proper CBFs and it is often hard to find them. In addition, Koller et al. (2018) proposed a learning-based model predictive control scheme that provides high-probability safety guarantees during the learning process under the assumption that both a dominant term of the state transition function and safe region are known a priori.

3 Problem Statement

We consider episodic finite-horizon CMDPs, which can be formally defined as a tuple

$$\mathcal{M} \coloneqq (\mathcal{S}, \mathcal{A}, P, T, r, g, s_1), \tag{1}$$

where $\mathcal{S}$ is a state space $\{s\}$, $\mathcal{A}$ is an action space $\{a\}$, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is an unknown, stochastic state transition function to map a state-action pair to a probability distribution over the next states, $T \in \mathbb{Z}_+$ is a fixed length of each episode, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a (bounded) reward function, $g : \mathcal{S} \times \mathcal{A} \to \{0, 1\}$ is an unknown binary safety function, and $s_1 \in \mathcal{S}$ is the initial state.[1] Crucially, in this paper, the safety feedback is provided as a *binary* value; that is, $g(s, a) = 1$ means that a state-action pair (s, a) is safe, and otherwise (s, a) is unsafe. At the time step t and the current state s_t, the agent takes the next action a_t, receiving the next state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$ as well as the safety observation $g(s_t, a_t)$, until the terminal time step T. We suppose that safety observations contain some independent zero-mean noise n_t. We assume that the noise n_t is sub-Gaussian with fixed (positive) parameters $\sigma \in \mathbb{R}_+$; that is, for all t, we have $\mathbb{E}[\, e^{\omega n_t} \mid \mathcal{G}_{t-1}\,] \leq e^{\omega^2\sigma^2/2}$, where $\{\mathcal{G}_t\}$ is increasing sequences of sigma fields such that n_t is $\mathcal{G}_t$-measurable with $\mathbb{E}[\, n_t \mid \mathcal{G}_{t-1}\,] = 0$. This assumption has been commonly made in previous work (e.g., Abbasi-yadkori, Pál, and Szepesvári (2011), Li, Lu, and Zhou (2017)).

A deterministic policy of an agent $\pi : \mathcal{S} \to \mathcal{A}$ represents a function to return actions. A metric of the quality of the policy π is the following value function, i.e., the expected value of cumulative rewards, which is defined as

$$V_t^\pi(s) \coloneqq \mathbb{E}_\pi \left[\sum_{\tau=t}^{T} r(s_\tau, a_\tau) \,\middle|\, s_t = s \right],$$

for all $s \in \mathcal{S}$ and $t \in [T]$, where the expectation $\mathbb{E}_\pi[\cdot]$ is taken over the trajectories $\{(s_\tau, a_\tau)\}_{\tau=t}^T$ induced by the policy π and true state transition dynamics P. We additionally define the following action-value function (i.e., Q-function) which means the expected value of total rewards when the agent starts from state-action pair (s, a) at step t and follows policy π, which is represented as

$$Q_t^\pi(s, a) \coloneqq \mathbb{E}_\pi \left[\sum_{\tau=t}^{T} r(s_\tau, a_\tau) \,\middle|\, s_t = s, a_t = a \right],$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $t \in [T]$.

A crucial point of this paper is that we wish the agent to take only *safe* actions at every time step t; that is, the agent needs to take a safe action a_t at a state s_t that satisfies the safety constraint; that is, $g(s_t, a_t) = 1$. As discussed in Section 1, however, at time t, the agent is required to execute safe actions in the long run so that there also will be future safe actions from time $t + 1$ to T. Hence, at every time step t, we impose the following safety requirement:

$$\Pr\Big\{ g(s_\tau, a_\tau) = 1 \quad \forall \tau \in [t, T] \Big\} \geq 1 - \delta, \tag{2}$$

[1] We assume that reward function is known and deterministic, but all results presented here extend to unknown stochastic cases.

where $\delta \in [0, 1]$ is a small positive scalar. Our safety constraint is probabilistic since it is extremely difficult to guarantee safety almost surely (i.e., probability of 1) due to the unknown stochastic state transition and safety functions.

Goal. Let us clearly describe the goal we wish to achieve in this paper. The objective of the agent is to obtain the optimal policy $\pi^\star : \mathcal{S} \to \mathcal{A}$ to maximize the value function $V_t^\pi(s_t)$ under the safety constraint (2) such that

$$\max_\pi V_t^\pi(s_t) \;\text{ s.t. }\; \Pr\Big\{ g(s_\tau, a_\tau) = 1 \;\; \forall \tau \in [t, T] \Big\} \geq 1 - \delta.$$

It is quite hard to guarantee the satisfaction of the aforementioned constraint. It is because even if the agent executes an action a_t at time t and state s_t such that

$$\Pr\Big\{ g(s_t, a_t) = 1 \Big\} \geq 1 - \delta, \tag{3}$$

there may *not* be any viable action at $s_{t+1} \sim P(s_t, a_t)$ and further future states. Thus, the agent must execute an action a_t to guarantee the constraint satisfaction not only for (s_t, a_t) but also for (s_τ, a_τ) for all $\tau \in [t+1, T]$. Our safety constraint (2) is challenging, which we will call the *long-term* safety constraint in the rest of this paper, while we will call (3) the *instantaneous* safety constraint.

Difficulties and assumptions. The aforementioned problem we wish to solve has several difficulties. First, if the binary safety function does not exhibit any regularity, it is impossible to infer the safety of state-action pairs. For example, if the safety function value is totally random, we can neither foresee danger nor guarantee safety. In addition, we suppose the state transition is stochastic and unknown a priori despite that the agent must guarantee the satisfaction of the long-term safety constraint. This difficulty requires us to explicitly incorporate the stochasticity of the state transition and its influence on future safety.

For the first difficulty, we assume that the safety function can be modeled as a GLM to deal with binary safety feedback. GLMs have been studied for sequential decision-making problems with binary feedback especially in (stateless) multi-armed bandit literature (Filippi et al. 2010; Li, Lu, and Zhou 2017; Faury et al. 2020) under the name of logistic bandits. Also, in (stateful) RL settings, Wachi, Wei, and Sui (2021) addressed a safe RL problem where the safety function is subject to a GLM under the assumption that the state transition is a priori known and deterministic. We now make the following assumption of the GLM structure of the safety function.

Assumption 1. *There exists a known feature mapping function $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^m$, unknown coefficient vectors $\boldsymbol{w}^\star \in \mathbb{R}^m$, and a fixed, strictly increasing (inverse) link function $\mu : \mathbb{R} \to [0, 1]$ such that*

$$\mathbb{E}[\, g(s, a) \mid s, a\,] = \mu\big(f^\star(s, a)\big), \tag{4}$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $f^\star : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a linear predictor defined as

$$f^\star(s, a) \coloneqq \langle \boldsymbol{\phi}(s, a), \boldsymbol{w}^\star \rangle, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{5}$$

Without loss of generality, we further assume $\|\boldsymbol{\phi}(s, a)\|_2 \leq 1$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $\|\boldsymbol{w}^\star\|_2 \leq \sqrt{m}$.

In the case of the binary safety function, a suitable choice of the link function is $\mu(x) = \exp(x)/(1+\exp(x))$, leading to the logistic regression model. GLMs are more general models, and one can verify that linear and integer-valued functions are special cases of GLMs with $\mu(x) = x$ and $\mu(x) = \exp(x)$ leading to the linear regression model and the Poisson regression model, respectively; hence, our method can be extended to other problem settings than the binary safety function.

In addition to the boundedness assumption on the feature vectors and safety function values, we make the following assumption regarding the link function.

Assumption 2. *The link function μ is twice differentiable, and the first and second-order derivatives are respectively bounded. Also, the link function μ satisfies $\xi = \inf_{\|\boldsymbol{w}-\boldsymbol{w}^\star\|\leq 1, \|\boldsymbol{\phi}\|\leq 1} \dot{\mu}(\langle \boldsymbol{\phi}, \boldsymbol{w}\rangle) > 0$.*

By making Assumptions 1 and 2, it is possible to guarantee the satisfaction of the instantaneous safety constraint (3) at the current time step t if there are feasible actions, as conducted in Wachi, Wei, and Sui (2021). In this paper, however, the agent must guarantee safety until the terminal time step T (i.e., long-term safety constraint) under stochastic state transition, which requires us to make further assumptions. If the feature mapping function drastically changes with minor differences in state-action pairs, it is extremely difficult to continue to guarantee safety until T. Thus, we assume the regularity of the feature mapping function as a form of Lipschitz continuity, which is written as follows:

Assumption 3. *For all $s, \bar{s} \in \mathcal{S}$ and $a, \bar{a} \in \mathcal{A}$, the feature mapping function $\boldsymbol{\phi}(\cdot,\cdot)$ is Lipschitz continuous with a constant $L_\phi \in \mathbb{R}_+$; that is,*

$$\|\boldsymbol{\phi}(s,a) - \boldsymbol{\phi}(\bar{s},\bar{a})\|_2 \leq L_\phi \cdot d_{\mathcal{SA}}((s,a),(\bar{s},\bar{a})), \quad (6)$$

where $d_{\mathcal{SA}}(\cdot,\cdot)$ is a distance metric on $\mathcal{S}\times\mathcal{A}$. For ease of exposition, we assume that $d_{\mathcal{SA}}$ satisfies $d_{\mathcal{SA}}((s,a),(\bar{s},\bar{a})) = d_{\mathcal{S}}(s,\bar{s}) + d_{\mathcal{A}}(a,\bar{a})$.

Intuitively, this assumption implies that, for similar state-action pairs (s,a) and $(\bar{s},\bar{a})$, the features $\boldsymbol{\phi}(s,a)$ and $\boldsymbol{\phi}(\bar{s},\bar{a})$ also exhibit similar values. This assumption is related to the common assumption in RL literature as represented by Lipschitz MDP (Asadi, Misra, and Littman 2018; Ok, Proutiere, and Tranos 2018).

Similarly, at a current state s, if the next state $s' \sim P(\cdot \mid s,a)$ induced by an "insignificant" action a (that tries to maintain the status quo) is far from s, the safety may drastically changes. Hence, we assume the existence of a Lipschitz-continuous conservative policy $\pi^\sharp : \mathcal{S} \to \mathcal{A}$ to suppress the state transition distance within a certain value. We then assume that, as far as similar policies to the conservative policy are executed, the state-transition distance can be suppressed. Specifically, for any policy π, we assume that the (one-step) state transition from time t to $t+1$ is upper-bounded according to the divergence between the actions taken by π and $\pi^\sharp$.

Assumption 4. *Let $L_\sharp \in \mathbb{R}_+$ be a positive scalar. There exists a known $L_\sharp$-Lipschitz continuous policy $\pi^\sharp : \mathcal{S} \to \mathcal{A}$ such that, for any states $s, \bar{s} \in \mathcal{S}$,*

$$d_{\mathcal{A}}(\pi^\sharp(s) - \pi^\sharp(\bar{s})) \leq L_\sharp \cdot d_{\mathcal{S}}(s,\bar{s}). \quad (7)$$

Also, with a positive scalar $\eta \in \mathbb{R}_+$, for any policy $\pi : \mathcal{S} \to \mathcal{A}$, the following inequality holds for all $s \in \mathcal{S}$:

$$\max_{s'\sim P(\cdot|s,\pi(s))} d_{\mathcal{S}}(s,s') \leq \bar{d} + \eta \cdot d_{\mathcal{A}}(\pi(s),\pi^\sharp(s)).$$

Remark 1. *Assumption 4 implies that the conservative policy $\pi^\sharp$ keeps the amount of the (one-step) state transition within a certain distance $\bar{d} \in \mathbb{R}_+$; that is,*

$$\max_{s'\sim P(\cdot|s,\pi^\sharp(s))} d_{\mathcal{S}}(s,s') \leq \bar{d}.$$

In the case of stochastic policies, we can use Kantorovich distance $K(\cdot,\cdot)$ to define the Lipschitz continuity of a policy; that is, $K(\pi(\cdot \mid s), \pi(\cdot \mid \bar{s})) \leq L_\pi \cdot d_{\mathcal{S}}(s,\bar{s})$. Thus, the following theoretical analysis can be extended to stochastic policy settings. This assumption is valid in many physical systems (e.g., control-affine systems). Intuitively, when the policy π is similar to the conservative policy $\pi^\sharp$, the upper bound of the state transition is guaranteed to be small. Assumption 4 implies that when $\pi = \pi^\sharp$, the state transition distance is always less than or equal to $\bar{d}$. Also, as π becomes far from $\pi^\sharp$, the distance can be larger with respect to the term $\eta \cdot d_{\mathcal{A}}(\pi,\pi^\sharp)$. The existence of $\pi^\sharp$ is not restrictive in practice for a number of applications, and similar notions have been adopted in many existing studies under the name of the stable or telescoping policies (Lin et al. 2021; Tsukamoto, Chung, and Slotine 2021). For instance, with the autonomous vehicle, one may select $\pi^\sharp$ as the one to move it at a low constant speed, and π is optimized such that it can move faster under the safety constraint.

4 Characterizing Safety

Based on the problem settings and assumptions presented in Section 3, we now present how to guarantee long-term safety. Optimism and pessimism are essential notions in RL. Conventionally, being optimistic has been well-adopted in online RL literature under the name of optimism in the face of uncertainty principle (Strehl and Littman 2008; Auer and Ortner 2007). In contrast, pessimism is also significant when an RL agent is trained from offline data (Jin, Yang, and Wang 2021; Buckman, Gelada, and Bellemare 2020) or needs to satisfy safety constraints (Bura et al. 2022). A natural way to incorporate optimism and pessimism is to derive the upper and lower bounds of the functions of interest, which can be conducted in a way backed by theory.

This paper expresses the upper and lower bounds in two ways. The first is inferred by the GLMs. While the advantage of this approach is to provide accurate estimation once a larger amount of dataset has been collected, the uncertainty term tends to be loose in the early phase of the training. The second is based on Lipschitz continuity. In contrast to the GLM-based approach, this approach provides moderate bounds regardless of the amount of collected data, which is typically useful in the early phase of training. Thus, intuitively, we aim to continue to derive tight bounds by deriving them using the approach based on Lipschitz continuity in the early phase and that based on the GLMs in the later phase.

4.1 Confidence Intervals Inferred by GLMs

We first present how to obtain theoretically-guaranteed confidence bounds inferred by the GLMs. Hereinafter, let the design matrix be $W_n = \sum_{j=1}^{n} \boldsymbol{\phi}(s_j, a_j)\boldsymbol{\phi}(s_j, a_j)^\top$, where $n \in \mathbb{Z}_+$ is the total number of data. Also, the weighted L_2-norm of $\boldsymbol{\phi}$ associated with W_n^{-1} is given by $\|\boldsymbol{\phi}\|_{W_n^{-1}} \coloneqq \sqrt{\boldsymbol{\phi}^\top W_n^{-1} \boldsymbol{\phi}}$. Here, the maximum-likelihood estimators (MLE) denoted as $\hat{\boldsymbol{w}}$ is calculated by solving the following equation:

$$\sum_{j=1}^{n} \big(g(s_j, a_j) - \mu(\langle\, \boldsymbol{\phi}(s_j, a_j), \boldsymbol{w} \,\rangle)\big)\boldsymbol{\phi}(s_j, a_j) = 0,$$

Based on Li, Lu, and Zhou (2017), the following lemma regarding the confidence bounds on $f^\star$ holds.

Lemma 1. *Let $\Delta > 0$ be given and $\beta = \frac{3\sigma}{\xi}\sqrt{\log \frac{3}{\Delta}}$. Then, with a probability of at least $1 - \Delta$, the MLE satisfies*

$$|f^\star(s, a) - \langle \boldsymbol{\phi}(s, a), \hat{\boldsymbol{w}} \rangle| \leq \beta \cdot \|\boldsymbol{\phi}(s, a)\|_{W_n^{-1}},$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Therefore, at time t and state s_t, by choosing the next action a_t such that $\langle \boldsymbol{\phi}(s_t, a_t), \hat{\boldsymbol{w}} \rangle - \beta \cdot \|\boldsymbol{\phi}(s_t, a_t)\|_{W_n^{-1}} \geq z$, we can also guarantee the satisfaction of $f^\star(s_t, a_t) \geq z$ with high probability, where $z \in \mathbb{R}$ is a certain threshold.

4.2 Bounds by Lipschitz Continuity

We then present the upper and lower bounds inferred by the Lipschitz continuity. Let us first define an important variable $x_t \in \mathbb{R}_+$ called maximum divergence from the conservative policy (MDCP) such that

$$d_{\mathcal{A}}(\pi(s_t), \pi^\sharp(s_t)) \leq x_t, \quad \forall t \in [T]. \tag{8}$$

The MDCP indicates how far the action taken by π is from that by $\pi^\sharp$. Hereinafter, the summation of this new variable x_t plays a critical role when dealing with the long-term safety constraint, and thus we define $X_{t_1}^{t_2} \coloneqq \sum_{\tau=t_1}^{t_2} x_\tau$ for any time steps $t_1, t_2 \in [T]$ with $t_1 < t_2$. We have the following two lemmas regarding the (true) safety linear predictor $f^\star$. See Appendices A.2 and A.3 for the proofs.

Lemma 2. *Suppose the policy π satisfies* (8)*. Let L_1, L_2, and L_3 be constants that are respectively defined as*

$$L_1 \coloneqq \sqrt{m} \cdot L_\phi, L_2 \coloneqq \bar{L}_\sharp \cdot \bar{d}, L_3 \coloneqq 2 + \eta \bar{L}_\sharp.$$

Set $\bar{t} \coloneqq T - t$ and recall that $X_{t+1}^{T-1} \coloneqq \sum_{\tau=t+1}^{T-1} x_\tau$. Finally, with $x_{t:T} \coloneqq x_t, x_{t+1}, \ldots, x_T$, define

$$\mathcal{F}(t, x_{t:T}) \coloneqq L_1 \left\{ L_2 \bar{t} + (L_3 - 1)x_t + L_3 X_{t+1}^{T-1} + x_T \right\}.$$

Then, we have

$$|f^\star(s_T, \pi(s_T)) - f^\star(s_t, \pi(s_t))| \leq \mathcal{F}(t, x_{t:T}).$$

Intuitively, Lemma 2 characterizes the present-to-future difference in terms of $f^\star$, which provides us the lower bound of the future safety linear predictor.

Lemma 3. *Define $f^\sharp(s) \coloneqq f^\star(s, \pi^\sharp(s))$ for all $s \in \mathcal{S}$. Also, suppose the policy π satisfies* (8)*. Then, we have*

$$\left| f^\star(s_t, \pi(s_t)) - f^\sharp(s_1) \right| \leq L_1 \left\{ L_2\, t + L_3 X_2^{t-1} + x_t \right\}.$$

In contrast to Lemma 2, Lemma 3 characterizes the past-to-present difference in terms of $f^\star$. Using this lemma, we infer the lower bound of $f^\star$ at the current time step t.

4.3 Resulting Lower Bound of $f^\star$

As we discussed previously, it is a simple yet powerful way to use the safety lower bound for introducing pessimism in safe RL. Let $\ell : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denote a lower bound of the true safety linear predictor, $f^\star$. To obtain a tighter bound, this paper combines the two lower bounds presented in Section 4.1 and 4.2, respectively. Specifically, based on Lemma 1 and 3, we obtain the following tighter bound:

$$\ell(s_t, a_t) \coloneqq \max\big(\ell_{\text{GLM}}(s_t, a_t), \ell_{\text{Lipschitz}}(s_t, a_t)\big), \tag{9}$$

where $\ell_{\text{GLM}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $\ell_{\text{Lipschitz}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are pessimistic safety linear predictors inferred by GLM and Lipshitz continuity, which are respectively defined as

$$\ell_{\text{GLM}}(s_t, a_t) \coloneqq \langle \boldsymbol{\phi}(s_t, a_t), \hat{\boldsymbol{w}} \rangle - \beta \cdot \|\boldsymbol{\phi}(s_t, a_t)\|_{W_n^{-1}},$$
$$\ell_{\text{Lipschitz}}(s_t, a_t) \coloneqq f^\sharp(s_1) - L_1 \left\{ L_2\, t + L_3 X_1^{t-1} + x_t \right\}.$$

4.4 Long-term Safety Guarantee

We present theoretical results regarding the lower bound of the safety linear predictor $f^\star$, which leads to the long-term safety guarantee defined by (2). We now present a lemma regarding the safety linear predictor at time t.

Lemma 4. *At every time step $t \in [T]$, we have*

$$f^\star(s_t, a_t) \geq \ell(s_t, a_t) \tag{10}$$

with a probability of at least $1 - \Delta$.

This lemma implies that we can guarantee the instantaneous safety constraint (3) by choosing the next action such that

$$\ell(s_t, a_t) \geq z, \quad \text{with} \quad \mu(z) = 1 - \delta, \tag{11}$$

with a probability of at least $1 - \Delta$.

In this paper, however, we need to additionally require the satisfaction of the long-term safety constraint; thus, we are particularly interested in future safety. We now provide the following lemma in terms of the pessimistic safety linear predictor at the terminal time step T:

Lemma 5. *Recall T is the terminal time step and set $\bar{t} = T - t$. At every time step $t \in [T]$, we have*

$$f^\star(s_T, a_T) \geq \ell(s_t, a_t) - \mathcal{F}(t, x_{t:T}),$$

with a probability of at least $1 - \Delta$.

Corollary 1. *Suppose, at state s_t, the agent with a policy π executes the action a_t while tuning $x_t, x_{t+1}, \ldots, x_T$ so that*

$$\ell(s_t, a_t) - \mathcal{F}(t, x_{t:T}) \geq z \tag{12}$$

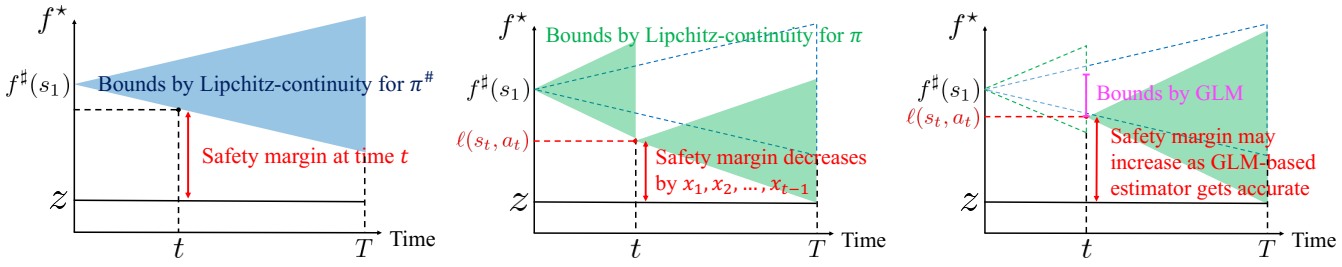
holds. Then, for all $\tau \in [t, T]$, there exist safe state-action pairs (s_τ, a_τ) such that:

$$f^\star(s_\tau, a_\tau) \geq z, \quad \tau \in [t, T], \tag{13}$$

with a probability of at least $1 - \Delta$.

The proofs of Lemma 5 and Corollary 1 are written in Appendix A.4.

Finally, we present a main theorem on the long-term safety constraint. Specifically, we guarantee that an agent continues to take safe actions from time t to T with a higher probability than a predefined threshold, by properly tuning the MDCPs, x_τ for all $\tau \in [t, T]$.



(a) Conservative policy (b) Early phase (Lipschitz) (c) Later phase (GLM)

Figure 2: (a) Bounds by Lipschitz continuity for the conservative policy. (b) In the early phase of training, the lower bound of the safety linear predictor at time t is typically characterized by the Lipschitz continuity, which decreases depending on the $x_1, x_2, \ldots, x_{t-1}$. Depending on the safety margin at time t, we need to control $x_t, x_{t+1}, \ldots, x_T$ for ensuring future safety. (c) As the training proceeds, the lower bound of the safety linear predictor can potentially be characterized by the GLMs, and the safety margin may increase.

Theorem 1. *Suppose, at state s_t, the agent executes the action a_t while tuning the MDCPs $x_t, x_{t+1}, \ldots, x_T$ so that (12) holds. Set $\delta := 1 - (1 - \mu(z))^{\bar{t}}$. Then, we have*

$$\Pr\Big\{g(s_\tau, a_\tau) = 1 \ \ \forall \tau \in [t, T]\Big\} \geq 1 - \delta, \quad \forall t \in [T],$$

— i.e. the long-term safety constraint is satisfied — with a probability of at least $1 - \Delta$.

This theorem guarantees that at every time step t, the agent can take safe actions from t to T with high probability, despite unknown, stochastic state transition and binary safety feedback. The proof sketch is as follows. By Corollary 1, when (12) is satisfied, $f^\star(x_\tau, a_\tau) \geq z$ holds for all $\tau \in [t, T]$ with high probability; that is, the existence of future safe actions are guaranteed with high probability. Theorem 1 provides a stricter safety guarantee than the one in existing safe RL literature with instantaneous safety constraints such as Wachi, Wei, and Sui (2021). If we tried to guarantee safety while using the instantaneous constraint (3), the agent would fall into worse situations and then lose the choices of safe actions due to the stochastic state transition.

5 LoBiSaRL Algorithm

We finally propose our LoBiSaRL algorithm. The algorithm flow is shown in Algorithm 1. Based on Theorem 1, we should solve the following policy optimization problem under a (conservative) long-term safety constraint:

$$\max_\pi V_t^\pi(s_t) \quad \text{subject to} \quad \ell(s_t, a_t) - \mathcal{F}(t, x_{t:T}) \geq z.$$

Note that, the term $\mathcal{F}(t, x_{t:T})$ can be transformed into

$$\mathcal{F}(t, x_{t:T}) = L_1 \cdot \Big\{ \underbrace{L_2 \bar{t}}_{\text{(A)}} + \underbrace{(L_3 - 1) x_t}_{\text{(B)}} + \underbrace{L_3 X_{t+1}^{T-1} + x_T}_{\text{(C)}} \Big\}.$$

The above inequality can be interpreted as follows. (A) is an inevitable term that even the conservative policy $\pi^\sharp$ cannot avoid. (B) depends only on the current action at time t, and (C) depends on the future actions from time $t+1$ to T. We can make (B) and (C) terms zero by executing the same actions as the conservative policy.

Algorithm 1: Long-term Binary Safe RL (LoBiSaRL)

Input: Initial Lagrange multiplier λ_1. Constants L_1, L_2, and L_3. conservative policy $\pi^\sharp$.
for iteration $i = 1, 2, \ldots$ **do**
  for time $t = 1, 2, \ldots, T$ **do**
    $\pi_t \leftarrow \operatorname{argmax}_\pi V_t^\pi(s_t) - \lambda_i \left(-x_t + L_3 X_t^{T-1} + x_T\right)$
    $A_t \leftarrow \{a \in \mathcal{A} \mid \ell(s_t, a) - L_1\{L_2 \bar{t} + (L_3 - 1) x_t\} \geq z\}$
    if $\pi_t(s_t) \in A_t$ **then**
      $a_t \leftarrow \pi_t(s_t)$
    else
      $a_t \leftarrow \operatorname{argmin}_{a \in A_t} \|a - \pi_t(s_t)\|_2$
    Take a_t and then receive a next state $s_{t+1} \sim P(s_t, a_t)$, reward $r(s, a)$, and (binary) safety $g(s, a)$.
    Update value function V_t^π
  $H_i := \min_t \left(\ell(s_t, \pi(s_t)) - z\right)$
  Update the Lagrange multiplier to λ_{i+1} based on H_i

A key to solving the aforementioned constrained policy optimization problem is how we tune x_τ for all $\tau = [t, T]$. Intuitively, we want to set x to be large in terms of reward maximization while x should be small in terms of long-term safety guarantee. Hence, we use a Lagrangian method to simultaneously maximize the expected cumulative reward while tuning the magnitude of x for the satisfaction of the safety constraint. Specifically, with a Lagrange multiplier $\lambda \in \mathbb{R}_+$, we solve the following max-min problem:

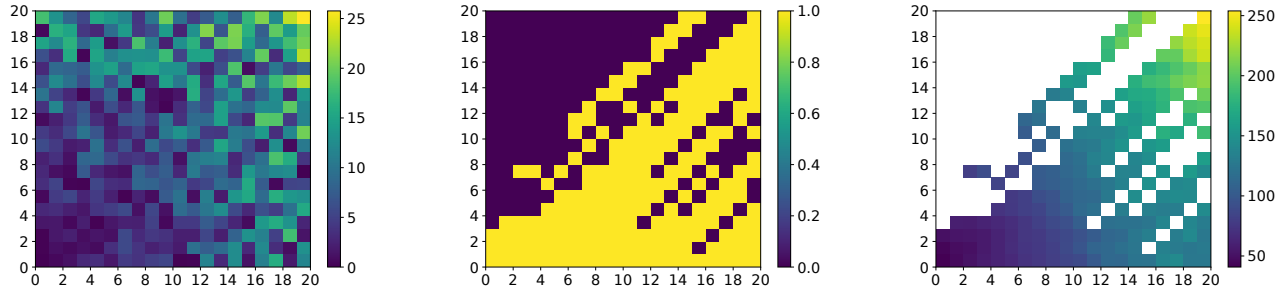
$$\max_\pi \min_{\lambda \geq 0} \ V_t^\pi(s_t) - \lambda \cdot (-x_t + L_3 X_t^{T-1} + x_T). \quad (14)$$

By setting λ large, we enforce the agent to make x small and thus execute similar actions to the conservative policy. When the conservative policy has much safety margin, the agent should explore the state and action spaces while taking more different actions. The degree of freedom is optimized by means of the Lagrange multiplier λ.

For the current policy π, the minimum safety requirement that the agent needs to satisfy at every time step t is

$$\ell(s_t, \pi(s_t)) - L_1 \left\{L_2 \bar{t} + (L_3 - 1) x_t\right\} \geq z. \quad (15)$$

The aforementioned inequality is derived by setting the (C) term to be 0, which corresponds to executing the same actions to the conservative policy from time $t+1$ to T. In other



(a) Reward function. (b) Binary safety function. (c) Value function with safety.

Figure 3: Example reward, binary safety, and value functions. In this paper, we consider a safe RL problem with binary safety feedback; thus, there is an unsafe region (the white region in the (c)) where the agent is not allowed to visit.

words, at time step t, the next action a_t must be chosen with the following "safe" action set:

$$A_t := \{a \in \mathcal{A} \mid \ell(s_t, a) - L_1\{L_2\bar{t} + (L_3 - 1)x_t\} \geq z\}.$$

To optimize the Lagrange multiplier λ, we define the following minimum safety margin at i-th episode:

$$H_i := \min_t \left(\ell(s_t, \pi(s_t)) - z\right).$$

When H_i is large, the agent is allowed to explore further by taking different actions from the conservative policy. In contrast, when H_i is small, the agent needs to prioritize safety without diverging from the conservative policy.

6 Experiments

In this section, we evaluate the performance of LoBiSaRL in a synthetic grid-world environment.

Settings. This environment is 20×20 square grids in which reward and safety functions are randomly generated. To avoid trivial situations where the optimal policy wanders around the initial position $(0, 0)$, we generate the reward function so that the reward-rich region is far from the initial state. The safety function is generated so as to follow a GLM, and the agent receives the binary safety feedback. At every time step, the agent takes an action from four action candidates (up, right, down, left). Also, the state transition function is stochastic; thus, the agent can go in the intended direction 80% of the time (if there is no wall). We provide 10 initial samples for initializing the GLM and set $T = 50$.

Baselines. We compare the performance of LoBiSaRL with four baselines. The first baseline is called RANDOM agent, which randomly chooses the next action without any consideration of reward and safety. The second is a UNSAFE agent. This agent purely maximizes the cumulative reward while ignoring the safety issues. The third baseline is a LINEAR agent. This algorithm is based on Amani, Thrampoulidis, and Yang (2021) to model the safety function via a linear model. The final baseline is a INSTANTANEOUS agent. This algorithm only considers the instantaneous safety constraint (3) as in Wachi, Wei, and Sui (2021) and cannot guarantee the satisfaction of the long-term constraint in an environment with the stochastic state transition.

	Reward	Unsafe actions
RANDOM	0.32 ± 0.24	23.2 ± 10.3
UNSAFE	$\mathbf{1.00 \pm 0.00}$	26.8 ± 13.6
LINEAR	0.73 ± 0.13	18.3 ± 5.7
INSTANTANEOUS	0.86 ± 0.10	3.3 ± 2.2
LoBiSaRL (Ours)	0.76 ± 0.12	$\mathbf{0.0 \pm 0.0}$

Table 2: Experimental results. Reward is normalized with respect to UNSAFE agent.

Results. Table 2 summarizes our experimental results. To obtain the results, we run each algorithm while generating 100 different random environments. As for safety, LoBiSaRL is the only algorithm to guarantee the satisfaction of the safety constraint in the long run. RANDOM, UNSAFE, and LINEAR execute a lot of unsafe actions. INSTANTANEOUS agent is much safer than the above three baselines but sometimes violates the safety constraint due to the stochasticity of the environment. In contrast, LoBiSaRL is often too conservative and the performance in terms of reward is worse than INSTANTANEOUS. Given that LoBiSaRL is an algorithm for safety-critical applications, however, it would be more important to guarantee long-term safety if the performance degradation is minor in terms of reward.

7 Conclusion

We formulate a safe RL problem with stochastic state transition and binary safety feedback and then propose an algorithm called LoBiSaRL. This algorithm maximizes the expected cumulative reward while guaranteeing the satisfaction of the long-term safety constraint. Under the assumptions regarding the Lipschitz continuity of the feature mapping function and the existence of a conservative policy, LoBiSaRL optimizes a policy while ensuring that there is at least one viable action until the terminal time step. We theoretically guarantee long-term safety and empirically evaluate the performance of LoBiSaRL comparing with several baselines. Moving forward, it is an interesting direction to improve performance in terms of reward.

References

Abbasi-yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved Algorithms for Linear Stochastic Bandits. In *Neural Information Processing Systems (NeurIPS)*.

Achiam, J.; Held, D.; Tamar, A.; and Abbeel, P. 2017. Constrained policy optimization. In *International Conference on Machine Learning (ICML)*.

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

Amani, S.; Thrampoulidis, C.; and Yang, L. 2021. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning (ICML)*.

Ames, A. D.; Coogan, S.; Egerstedt, M.; Notomista, G.; Sreenath, K.; and Tabuada, P. 2019. Control barrier functions: Theory and applications. In *European control conference (ECC)*.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Asadi, K.; Misra, D.; and Littman, M. 2018. Lipschitz continuity in model-based reinforcement learning. In *International Conference on Machine Learning (ICML)*.

Auer, P.; and Ortner, R. 2007. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*.

Bennett, A.; Misra, D.; and Kallus, N. 2023. Provable Safe Reinforcement Learning with Binary Feedback. In *International Conference on Artificial Intelligence and Statistics (AISTAT)*.

Buckman, J.; Gelada, C.; and Bellemare, M. G. 2020. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*.

Bura, A.; HasanzadeZonuzy, A.; Kalathil, D.; Shakkottai, S.; and Chamberland, J.-F. 2022. DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *Neural Information Processing Systems (NeurIPS)*.

Cheng, R.; Orosz, G.; Murray, R. M.; and Burdick, J. W. 2019. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *AAAI conference on artificial intelligence (AAAI)*.

Chow, Y.; Ghavamzadeh, M.; Janson, L.; and Pavone, M. 2017. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research (JMLR)*, 18(1): 6070–6120.

Dulac-Arnold, G.; Levine, N.; Mankowitz, D. J.; Li, J.; Paduraru, C.; Gowal, S.; and Hester, T. 2021. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 1–50.

Faury, L.; Abeille, M.; Calauzènes, C.; and Fercoq, O. 2020. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning (ICML)*.

Filippi, S.; Cappe, O.; Garivier, A.; and Szepesvári, C. 2010. Parametric bandits: The generalized linear case. *Neural Information Processing Systems (NeurIPS)*.

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

Jin, Y.; Yang, Z.; and Wang, Z. 2021. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*.

Koller, T.; Berkenkamp, F.; Turchetta, M.; and Krause, A. 2018. Learning-based model predictive control for safe exploration. In *IEEE conference on decision and control (CDC)*.

Li, L.; Lu, Y.; and Zhou, D. 2017. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning (ICML)*.

Lin, Y.; Hu, Y.; Shi, G.; Sun, H.; Qu, G.; and Wierman, A. 2021. Perturbation-based regret analysis of predictive control in linear time varying systems. *Advances in Neural Information Processing Systems*, 34: 5174–5185.

Ok, J.; Proutiere, A.; and Tranos, D. 2018. Exploration in structured reinforcement learning. *Neural Information Processing Systems (NeurIPS)*, 31.

Rockafellar, R. T.; Uryasev, S.; et al. 2000. Optimization of conditional value-at-risk. *Journal of risk*, 2: 21–42.

Roderick, M.; Nagarajan, V.; and Kolter, Z. 2021. Provably safe PAC-MDP exploration using analogies. In *International Conference on Artificial Intelligence and Statistics (AISTAT)*.

Stooke, A.; Achiam, J.; and Abbeel, P. 2020. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning (ICML)*.

Strehl, A. L.; and Littman, M. L. 2008. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8): 1309–1331.

Tessler, C.; Mankowitz, D. J.; and Mannor, S. 2018. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*.

Tsukamoto, H.; Chung, S.-J.; and Slotine, J.-J. E. 2021. Contraction theory for nonlinear stability analysis and learning-based control: A tutorial overview. *Annual Reviews in Control*, 52: 135–169.

Turchetta, M.; Berkenkamp, F.; and Krause, A. 2016. Safe exploration in finite Markov decision processes with Gaussian processes. In *Neural Information Processing Systems (NeurIPS)*.

Wachi, A.; and Sui, Y. 2020. Safe reinforcement learning in constrained Markov decision processes. In *International Conference on Machine Learning (ICML)*.

Wachi, A.; Wei, Y.; and Sui, Y. 2021. Safe Policy Optimization with Local Generalized Linear Function Approximations. *Neural Information Processing Systems (NeurIPS)*.

Yang, L.; and Wang, M. 2020. Reinforcement Learning in Feature Space: Matrix Bandit, Kernels, and Regret Bound. In *International Conference on Machine Learning (ICML)*.