Safe Exploration in Reinforcement Learning: A Generalized Formulation and Algorithms

Akifumi Wachi
LINE Corporation
akifumi.wachi@linecorp.com

Wataru Hashimoto
Osaka University
hashimoto@is.eei.eng.osaka-u.ac.jp

Xun Shen
Osaka University
shenxun@eei.eng.osaka-u.ac.jp

Kazumune Hashimoto
Osaka University
hashimoto@eei.eng.osaka-u.ac.jp

Abstract

Safe exploration is essential for the practical use of reinforcement learning (RL) in many real-world scenarios. In this paper, we present a generalized safe exploration (GSE) problem as a unified formulation of common safe exploration problems. We then propose a solution of the GSE problem in the form of a meta-algorithm for safe exploration, MASE, which combines an unconstrained RL algorithm with an uncertainty quantifier to guarantee safety in the current episode while properly penalizing unsafe explorations *before actual safety violation* to discourage them in future episodes. The advantage of MASE is that we can optimize a policy while guaranteeing with a high probability that no safety constraint will be violated under proper assumptions. Specifically, we present two variants of MASE with different constructions of the uncertainty quantifier: one based on generalized linear models with theoretical guarantees of safety and near-optimality, and another that combines a Gaussian process to ensure safety with a deep RL algorithm to maximize the reward. Finally, we demonstrate that our proposed algorithm achieves better performance than state-of-the-art algorithms on grid-world and Safety Gym benchmarks without violating any safety constraints, even during training.

1 Introduction

Safe reinforcement learning (RL) is a promising paradigm that enables policy optimizations for safety-critical decision-making problems (e.g., autonomous driving, healthcare, and robotics), where it is necessary to incorporate safety requirements to prevent RL policies from posing risks to humans or objects [14]. As a result, safe exploration has received significant attention in recent years as a crucial issue for ensuring the safety of RL during both the learning and execution phases [6].

Safe exploration in RL has typically been addressed by formulating a constrained RL problem in which the policy optimization is subject to safety constraints [9, 18]. While there have been many attempts under different types of constraint representations (e.g., expected cumulative cost [3], CVaR [29]), satisfying constraints almost surely or with high probability received less attention to date. Imagine safety-critical applications such as planetary exploration where even a single constraint violation may result in catastrophic failure. NASA's engineers hope Mars rovers to ensure safety at least with high probability [8]; thus, constraint satisfaction "on average" does not fit their purpose.

While several algorithms have addressed this problem with this stricter notion of safety, there are several formulations in terms of how the constraints are represented, including cumulative [32], state [39], and instantaneous constraints [41], which respectively correspond to Problems 1, 2, and 3

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

as we will discuss shortly in Section 2. Unfortunately, there has been limited discussion on the relationships between these approaches, making it challenging for researchers to acquire a systematic understanding of the field as a whole. If a generalized problem were to be formulated, then the research community could pool their efforts to develop suitable algorithms.

A closer examination of existing algorithms that span the entire theory-to-practice spectrum reveals several areas for improvement. Practical algorithms using deep RL (e.g., [32],[39],[40]) may provide satisfactory performance after convergence, but do not usually guarantee safety during training. In contrast, theoretical studies (e.g., [4], [41]) that guarantee safety with high probability during training often have limitations, such as relying on strong assumptions (e.g., known state transition) or experiencing decreased performance in complex environments. In summary, many algorithms have been proposed in various safe RL formulations, but the creation of a safe exploration algorithm that is both practically useful and supported by theoretical foundations remains an open problem.

Contributions. We first present a generalized safe exploration (GSE) problem and prove its generality compared with existing safe exploration problems. By taking advantage of the tractable form of the safety constraint in the GSE problem, we establish a meta-algorithm for safe exploration, MASE. This algorithm employs an uncertainty quantifier for a high-probability guarantee that the safety constraints are not violated and penalizes the agent *before* safety violation, under the assumption that the agent has access to an "emergency stop" authority. Our MASE is both practically useful and theoretically well-founded, which allows us to optimize a policy via an arbitrary RL algorithm under the high-probability safety guarantee, even during training. We then provide two specific variants of MASE with different uncertainty quantifiers. One is based on generalized linear models (GLMs), for which we theoretically provide high-probability guarantees of safety and near-optimality. The other is more practical, combining a Gaussian process (GP, [27]) to ensure safety with a deep RL algorithm to maximize the reward. Finally, we show that MASE performs better than state-of-the-art algorithms on the grid-world and Safety Gym [28] without violating any safety constraints, even during training.

2 Preliminaries

Definitions. We consider an episodic safe RL problem in a constrained Markov decision process (CMDP, [3]), $M = \langle S, A, H, P, r, g, s_1 \rangle$, where S is a state space, A is an action space, $H \in \mathbb{Z}_{>0}$ is a (fixed) length of each episode, $P : S \times A \times S \to [0, 1]$ is a state transition probability, $r : S \times A \to [0, 1]$ is a reward function, $g : S \times A \to [0, 1]$ is a safety (cost) function, and $s_1 \in S$ is an initial state. At each discrete time step, with a given (fully-observable) state s, the agent selects an action a with respect to its policy $\pi : S \to A$, receiving the new state s', reward r, and safety cost g. Though we assume a deterministic policy, our core ideas can be extended to stochastic policy settings. Given a policy π, the value and action-value functions in a state s at time h are respectively defined as

$$V^{\pi}_{r,h}(s) := \mathbb{E}\left[\sum_{h'=h}^{H} \gamma_r^{h'} r(s_{h'}, a_{h'}) \mid s_h = s\right]$$

and $Q^{\pi}_{r,h}(s, a) := \mathbb{E}\left[\sum_{h'=h}^{H} \gamma_r^{h'} r(s_{h'}, a_{h'}) \mid s_h = s, a_h = a\right]$, where the expectation $\mathbb{E}$ is taken over the random state-action sequence $\{(s_{h'}, a_{h'})\}_{h'=h}^{H}$ induced by the policy π. Additionally, $\gamma_r \in (0, 1]$ is a discount factor for the reward function. In the remainder of this paper, we define $V_{\max} := \frac{1}{1-\gamma_r}$ and let $\mathcal{T}_h : (S \times A \to \mathbb{R}) \to (S \times A \to \mathbb{R})$ denote the Bellman update operator $\mathcal{T}_h(Q)(s, a) := \mathbb{E}[r(s_h, a_h) + \gamma_r V_Q(s_{h+1}) \mid s_h = s, a_h = a]$, where $V_Q(s) := \max_{a \in A} Q(s, a)$.

Three common safe RL problems. We tackle safe RL problems where constraints must be satisfied almost surely, even during training. While such problems have garnered attention in the research community, there are several types of formulations, and their relations are yet to be fully investigated.

One of the most popular formulations for safe RL problems involves maximizing $V^{\pi}_r := V^{\pi}_{r,1}(s_1)$ under the constraint that the cumulative cost is less than a threshold, which is described as follows:

Problem 1 (Almost surely safe RL with cumulative constraint [32])**.**

$$\max_{\pi} V^{\pi}_r \quad \text{subject to} \quad \Pr\left[\sum_{h=1}^{\infty} \gamma_g^{h} g(s_h, a_h) \le \alpha_1\right] = 1,$$

where $\alpha_1 \in \mathbb{R}_{\ge 0}$ is a constant representing a threshold, and $\gamma_g \in (0, 1]$ is a discount factor for g.

Observe that, the expectation is not taken regarding the safety constraint in Problem 1. This problem was studied in [32], which is stricter than the conventional one where the expectation is taken with respect to the cumulative safety cost function (i.e., $\mathbb{E}[\sum_{h=1}^{H} \gamma^h g(s_h, a_h)] \leq \alpha_1$).

Another popular formulation involves leveraging the state constraints so that safety corresponds to avoiding visits to a set of unsafe states. This type of formulation has been widely adopted by previous studies on safe-critical robotics tasks [38–40, 45], which is written as follows:

Problem 2 (Safe RL with state constraints)

$$\max V_r \quad \text{subject to} \quad \mathbb{E}\left[\sum_{h=1}^{H} \gamma^h \mathbb{I}(s_h \in \mathcal{S}_{\text{unsafe}})\right] \leq \alpha_2,$$

where $\alpha_2 \in \mathbb{R}_{\geq 0}$ is a threshold, $\mathbb{I}(\cdot)$ is the indicator function, and $\mathcal{S}_{\text{unsafe}} \subseteq \mathcal{S}$ is a set of unsafe states.

Finally, some existing studies formulate safe RL problems via an instantaneous constraint, attempting to ensure safety even during the learning stage while aiming for extremely safety-critical applications such as planetary exploration [42] or healthcare [36]. Such studies typically require the agent to satisfy the following instantaneous safety constraint at every time step.

Problem 3 (Safe RL with instantaneous constraints)

$$\max V_r \quad \text{subject to} \quad \Pr\left[g(s_h, a_h) \leq \alpha_3 \mid P, \pi\right] = 1, \quad \forall h \in [1, H],$$

where $\alpha_3 \in \mathbb{R}_{\geq 0}$ is a time-invariant safety threshold.

3 Problem Formulation

This paper also requires an agent to optimize a policy under a safety constraint, as in the three common safe RL problems. We seek to find the optimal policy $\pi^\star : \mathcal{S} \to \mathcal{A}$ of the following problem, which will hereinafter be referred to as the "generalized" safe exploration (GSE) problem:

Problem 4 (GSE problem) Let $b_h \in \mathbb{R}$ denote a time-varying threshold.

$$\max V_r \quad \text{subject to} \quad \Pr\left[g(s_h, a_h) \leq b_h \mid P, \pi\right] = 1, \quad \forall h \in [1, H].$$

This constraint is instantaneous, which requires the agent to learn a policy without a single constraint violation not only after convergence but also during training. We assume that the threshold is myopically known; that is, b_h is known at time h, but unknown before that. Crucially, at every time step h, since s_h is a fully observable state and the agent's policy is deterministic, we will use a simplified inequality represented as $g(s_h, a_h) \leq b_h$ in the rest of this paper. This constraint is akin to that in Problem 3, with the difference that the safety threshold is time-varying.

Importance of the GSE problem. Though our problem may not seem relevant to Problems 1 and 2, we will shortly present and prove a theorem on the relationship between the GSE problem and the three common safe RL problems.

Theorem 3.1. Problems 1, 2, and 3 can be transformed into the GSE problem (i.e., Problem 4).

See Appendix B for the proof. In other words, the feasible policy space in the GSE problem can be identical to those in the other three problems by properly defining the safety cost function g and threshold b_h. Crucially, Problem 1 is a special case of the GSE problem with $b_h = \alpha_h$ for all h, where $\alpha_{h+1} = \alpha_h - g(s_h, a_h)$ with $\alpha_0 = \alpha_1$. It is particularly beneficial to convert Problems 1 and 2, which have additive constraint structures, to the GSE problem, which has an instantaneous constraint. The accurate estimation of the cumulative safety value in Problems 1 and 2 is difficult because they depend on the trajectories induced by a policy. Dealing with the instantaneous constraint in the GSE problem is easier, both theoretically and empirically. Also, especially when the environment is time-varying (e.g., there are moving obstacles), the GSE problem is more useful than Problem 3.

Typical CMDP formulations with expected cumulative (safety) cost are out of the scope of the GSE problem. In such problems, the safety notion is milder; hence, although many advanced deep RL algorithms have been actively proposed that perform well in complex environments after convergence,

their performance in terms of safety during learning is usually low, as reported by Stooke et al. [38] and Wachi et al. [43]. Risk-constrained MDPs are also important safe RL problems that are not covered by the GSE problem; they have been widely studied by representing risk as a constraint on some conditional value-at-risk [11] or using chance constraints [24, 26].[1]

Difficulties and Assumptions. Theorem 3.1 insists that the GSE problem covers a wide range of safe RL formulations and is worth solving, but the problem is intractable without assumptions. We now discuss the difficulties in solving the GSE problem, and then list the assumptions in this paper.

The biggest difficulty with the GSE problem lies in the fact that there may be no viable safe action given the current state s_h, safety cost g, and threshold b_h. When $b_h = 0.1$ and $g(s_h, a) = 0.5, \forall a \in \mathcal{A}$, the agent has no viable action for ensuring safety. The agent needs to guarantee safety, even during training, where little environmental information is available; hence, it is significant for the agent to avoid such situations where there is no action that guarantees safety. Another difficulty is related to the regularity of the safety cost function and the strictness of the safety constraint. In this paper, the safety cost function is unknown a priori.; thus, when the safety cost does not exhibit any regularity, the agent can neither infer the safety of decisions nor guarantee safety almost surely.

To address the first difficulty mentioned above, we use Assumptions 3.2 and 3.3.

Assumption 3.2 (Safety margin) There exists $\tau \in \mathbb{R}_{>0}$ such that $\Pr[g(s_h, a_h) \leq b_h \mid P, \tau] = 1, \forall h \in [1, H]$.

Assumption 3.3 (Emergency stop action) Let a_E be an emergency stop action such that $P(s_{\text{E}} \mid s, a_E) = 1$ for all $s \in \mathcal{S}$. The agent is allowed to execute the emergency stop action and reset the environment if and only if the probability of guaranteed safety is not sufficiently high.

Assumption 3.2 is mild; this is similar to the Slater condition, which is widely adopted in the CMDP literature [13, 25]. We consider Assumption 3.3 is also natural for safety-critical applications because it is usually better to guarantee safety, even with human interventions, if the agent requires help in emergency cases. In some applications (e.g., the agent is in a hazardous environment), however, emergency stop actions should often be avoided because of the expensive cost of human intervention. In such cases, the agent needs to learn a reset policy allowing them to return to the initial state as in Eysenbach et al. [15], rather than asking for human help, which we will leave to future work.

As for the second difficulty, we assume that the safety cost function belongs to a class where uncertainty can be estimated and guarantee the satisfaction of the safety constraint with high probability. We present an assumption regarding an important notion called an uncertainty quantifier:

Assumption 3.4 (ϵ-uncertainty quantifier) Let $\mu : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denote the estimated mean function of safety. There exists an uncertainty quantifier $\epsilon : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that $|g(s, a) - \mu(s, a)| \leq \epsilon(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, with a probability of at least $1 - \delta$.

4 Method

We propose MASE for the GSE problem, which combines an unconstrained RL algorithm with additional mechanisms for addressing the safety constraints. The pseudo-code is provided in Algorithm 1, and a conceptual illustration can be seen in Figure 1.

The most notable feature of MASE is that safety is guaranteed via the uncertainty quantifier and the emergency stop action (lines 2 – 9). The ϵ-uncertainty quantifier is particularly useful because we can guarantee that the confidence bound contains the true safety cost function, that is, $g(s, a) \in [\mu(s, a) - \epsilon(s, a), \mu(s, a) + \epsilon(s, a)]$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. This means that, if the agent chooses actions such that $\mu(s_h, a_h) + \epsilon(s_h, a_h) \leq b_h$, then $g(s_h, a_h) \leq b_h$ holds with a probability of at least $1 - \delta$. Regarding the first difficulty mentioned in Section 3, it is crucial that there is at least one safe action. Thus, at every time step h, the agent computes a set of actions that are considered to satisfy the safety constraints with a probability at least $1 - \delta$ given the state s_h and threshold b_h. This is represented as

$$\mathcal{A}_h^+ := \{ a \in \mathcal{A} \mid \min\{1, \mu(s_h, a) + \epsilon(s_h, a)\} \leq b_h \}.$$

Whenever the agent identifies that at least one action will guarantee safety, the agent is required to choose an action within $\mathcal{A}_h^+$ (line 3). The emergency stop action is executed if and only if there is no

[1] The solution in the GSE problem is guaranteed to be a conservative approximation of that in safe RL problems with chance constraints. For more details, see Appendix C.

Algorithm 1 Meta-Algorithm for Safe Exploration (MASE)

```
for episode t = 1, 2, ..., T do
    for time h = 1, 2, ..., H do
        Take "safe" action a_h = π(s_h) within A_h^+                    . Execute only safe actions
        Receive reward r(s_h, a_h), safety cost c(s_h, a_h), and next state s_{h+1}
        Update safety threshold d_{h+1}
        if A_{h+1}^+ = ∅ then
            Compute b(s_h, a_h) = r(s_h, a_h) + c / (min_{a∈A} c(s_{h+1}, a))    . Penalty for the emergency stop action
            Append (s_h, a_h, b(s_h, a_h), s_{h+1}) to D
            break (i.e., take action a_b)                               . Execute the emergency stop action
        else
            Append (s_h, a_h, r(s_h, a_h), s_{h+1}) to D
    Optimize a policy π based on D via an (unconstrained) RL algorithm
    Update the uncertainty quantifier and rewrite D
```

viable action satisfying the safety constraint (i.e., $A^+_{h+1} \neq \emptyset$); that is, the agent is allowed to execute a_b and start a new episode from an initial safe state (line 6 in Alg. 1). The safety cost is upper-bounded by 1 because $c \in [0, 1]$ by definition. Note that MASE proactively avoids unsafe actions by selecting the emergency stop action to take beforehand; this is in contrast to Sun et al. [37], whose method terminates the episode immediately after the agent has already violated a safety constraint.

When safety is guaranteed in the manner described above, the question remains as to how to obtain a policy that maximizes the expected cumulative reward. As such, we first convert the original CMDP M to the following unconstrained MDP

$$M' := \langle S, \{A, a_b\}, r, H, P, b, s_1 \rangle.$$

The changes from M lie in the action space and the reward function, as well as in the absence of the safety cost function. First, the action space is augmented so that the agent can execute the emergency stop action a_b. The second modification concerns the reward function. When executing the emergency stop action a_b, the agent is penalized as its sacrifice so that the same situation will not occur in future episodes; hence, we modify the reward function as follows:

$$b(s_h, a_h) = \begin{cases} c / \min_{a \in A} c(s_{h+1}, a) & \text{if } A^+_{h+1} = \emptyset, \\ r(s_h, a_h) & \text{otherwise,} \end{cases} \tag{1}$$

where $c \in \mathbb{R}_{>0}$ is a positive scalar representing a penalty for performing the emergency stop. This penalty is assigned to the state-action pair (s_h, a_h) that placed the agent into the undesirable situation at time step $h+1$, represented as $A^+_{h+1} = \emptyset$ (see Figure 1).

To show that MASE is a reasonable safe RL algorithm, we express the following intuitions. Consider the ideal situation in which the safety cost function is accurately estimated for any state-action pairs; that is, $c(s, a) = 0$ for all (s, a). In this case, all emergency stop actions are properly executed, and the safety constraint will be violated at the next time step if the agent executes other actions. It is reasonable for the agent to receive a penalty of $b(s, a) = 1$ because this state-action pair surely causes a safety violation without the emergency stop action. Unfortunately, however, the safety cost is uncertain and the agent conservatively executes a_b although there are still actions satisfying the safety constraint, especially in the early phase of training; hence, we increase or reduce the penalty according to the magnitude of uncertainty in (1) to avoid an excessively large penalty.

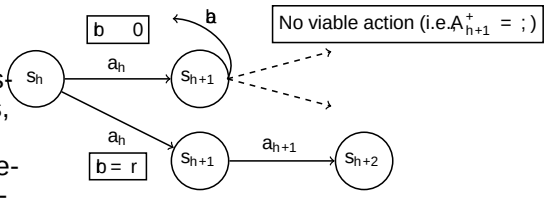


Figure 1: Conceptual illustration of MASE. At every time step h, the agent chooses action a_h within A^+_h. If there is no safe action at state s_{h+1} satisfying the constraint, the emergency stop action a_b is executed and the agent receives a large penalty for (s_h, a_h).

We must carefully consider the fact that the quality of information regarding the modified reward function b is uneven in the replay buffer D. Specifically, in the early phase of training, the

uncertainty quantifier is loose. Hence, the emergency stop action is likely to be executed even if viable actions remain; that is, the agent will receive unnecessary penalties. In contrast, the emergency stop actions in later phases are executed with confidence, as reflected by the tight uncertainty quantifier. Thus, as in line 15, we rewrite the replay buffer $\mathcal{D}$ while updating $\mathcal{E}$ depending on the model in terms of the safety cost function (as for specific methods to update $\mathcal{E}$, see Sections 5 and 6).

Connections to shielding methods. The notion of the emergency stop action is akin to shielding [2, 20] which has been actively studied in various problem settings including partially-observable environments [10] or multi-agent settings [23]. Thus, MASE can be regarded as a variant of shielding methods (especially, preemptive shielding [2]) that is specialized for the GSE problem. On the other hand, MASE does not only block unsafe actions but also provides proper penalties for executing the emergency stop actions based on the uncertainty quantifier, which leads to rigorous theoretical guarantees presented shortly. Such theoretical advantages can be enjoyed in many safe RL problems because of the wide applicability of the GSE problem backed by Theorem 3.1.

Advantages of MASE. Though certain existing algorithms for Problem 3 (i.e., the closest problem to the GSE problem) theoretically guarantee safety during learning, several strong assumptions are needed, such as a known and deterministic state transition and regular safety function [41] and a known feature mapping function that is linear with respect to transition kernels, reward, and safety as in [5]. Such algorithms have little affinity with deep RL; thus, their actual performance in complex environments tends to be poor. In contrast, MASE is compatible with any advanced RL algorithms, which can also handle various constraint formulations while maintaining the safety guarantee.

Validity of MASE. We conclude this section by presenting the following two theorems to show that our MASE produces reasonable operations in solving the GSE problem.

Theorem 4.1. Under Assumption 3.4, MASE guarantees safety with a probability of at least $1-\delta$.

Theorem 4.2. Assume that the safety cost function is estimated for any state-action pairs with an accuracy of better than $\frac{\eta}{2}$; that is, $|\hat{c}(s,a) - c(s,a)| \leq \frac{\eta}{2}$ for all (s,a). Set $c \in \mathbb{R}$ to be a sufficiently large scalar such that $c > \frac{V_{\max}}{2^{H}-r}$. Then, the optimal policy in $\mathcal{M}$ is identical to that in $\mathcal{M}$.

See Appendix D for the proofs. Unfortunately, obtaining an uncertainty quantifier that works in general cases is highly challenging. To develop a feasible model for the uncertainty quantification, we assume that the safety cost can be modeled via a GLM in Section 5 and via a GP in Section 6.

5 A Provable Algorithm under Generalized Linear CMDP Assumptions

In this section, we focus on CMDPs with generalized linear structures and analyze the theoretical properties of MASE. Specifically, we provide a provable algorithm to use a class of GLMs denoted as $\mathcal{F}$ for modeling $Q^{?}_{r,h} := Q_{r,h}$ and $g^{?}$, and then provide theoretical results on safety and optimality.

5.1 Generalized Linear CMDPs

We extend the assumption in Wang et al. [44] from unconstrained MDPs to CMDPs settings. Our assumption is based on GLMs as with [41] that makes a strictly weaker assumption than their Linear MDP assumption [19, 46]. As preliminaries, we first list the necessary definitions and assumptions.

Definition 5.1 (GLMs). Let $d \in \mathbb{Z}_{>0}$ be a feature dimension and $\mathcal{B}^d := \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ be the l_2 ball in $\mathbb{R}^d$. For a known feature mapping function: $\mathcal{S} \times \mathcal{A} \to \mathcal{B}^d$ and a known link function $f : [-1, 1] \to [-1, 1]$, the class of generalized linear model is denoted as $\mathcal{F} = \{f(\langle \cdot, \theta \rangle) : \theta \in \mathcal{B}^d\}$ where $\phi_{s,a} := \phi(s,a)$.

Assumption 5.2 (Regular link function). The link function $f(\cdot)$ is twice differentiable and is either monotonically increasing or decreasing. Furthermore, there exist absolute constants $0 < \underline{\kappa} < \bar{\kappa} < 1$ and $M < 1$ such that $\underline{\kappa} < |f'(x)| < \bar{\kappa}$ and $|f''(x)| < M$ for all $|x| \leq 1$.

This assumption on the regular link function is standard in previous studies [16, 21]. Linear and logistic models are the special cases of the GLM where the link functions are defined as $f(x) = x$ and $f(x) = 1/(1 + e^{-x})$. In both cases, the link functions satisfy Assumption 5.2.

We finally make the assumption of generalized linear CMDPs (GL-CMDPs), which extends the notion of the optimistic closure for unconstrained MDP settings in Wang et al. [44].

Assumption 5.3(GL-CMDP). For any $1 \le h < H$ and $u \in \mathcal{F}_{up}$, we have $\mathcal{T}_h(u) \in \mathcal{F}$ and $g \in \mathcal{F}$.

Recall that $\mathcal{T}_h$ is the Bellman update operator. In Assumption 5.3, with a positive semi-definite matrix $A \in \mathbb{R}^{d \times d}$ and a fixed positive constant $\kappa_{\max} \in \mathbb{R}_{>0}$, we define

$$\mathcal{F}_{up} := \{(s,a) \mapsto \min\{V_{max}, f(h_{s,a}, \theta) + \beta \|\phi_{s,a}\|_A\} : \theta \in B^d; 0 \le \beta \le \kappa_{\max}, \|A\|_{op} \le 1\},$$

where $\|x\|_A := \sqrt{x^\top A x}$ is the matrix Mahalanobis seminorm, and $\|A\|_{op}$ is the matrix operator norm. For simplicity, we suppose the same link functions for the Q-function and the safety cost function, but it is acceptable to use different link functions. Note that Assumption 5.3 is a more general assumption than Amani et al.[5] that assumes linear transition kernel, reward, and safety cost functions or Wachi et al. [43] that assumes a known transition and GLMs in terms of reward and safety cost functions.

5.2 GLM-MASE Algorithm

We introduce an algorithm GLM-MASE under Assumptions 5.2 and 5.3. Hereinafter, we explicitly denote the episode for each variable. For example, we let $s_h^{(t)}$ or $a_h^{(t)}$ denote a state or action at the time step h of episode t. We also let $e_h^{(t)} := (s_h^{(t)}, a_h^{(t)})$ for more concise notations.

Uncertainty quantifiers. To actualize MASE in the generalized linear CMDP settings, we first need to consider how to obtain the uncertainty quantifier in terms of the safety cost function. Since we assume $g \in \mathcal{F}$, we can define the -uncertainty quantifier based on the existing studies on GLMs, especially in the field of multi-armed bandits [22, 16]. Based on Assumptions 5.2 and 5.3, we now provide a lemma regarding the uncertainty quantifier on safety.

Lemma 5.4. Suppose Assumptions 5.2 and 5.3 hold. Set $\delta = \frac{1}{TH}$. With a universal constant $C \in \mathbb{R}_{>0}$, let $C_g := C\kappa^{-1}\sqrt{1 + M + \lambda + d^2 \ln\frac{1+\lambda+\kappa_{\max}}{\delta}}$. Define

$$\sigma_g(s,a) := C_g \|\phi_{s,a}\|_{\Lambda_{h,t}^{-1}} \quad \text{with} \quad \Lambda_{h,t} := \sum_{\tau} e_h^{(\tau)} e_h^{(\tau)\top} + I,$$

where I is the identity matrix. Let $\hat{\theta}^g_{h,t} \in \mathbb{R}^d$ be the ridge estimate, which is computed by $\hat{\theta}^g_{h,t} := \arg\min_{\|\theta\|_2 \le 1} \sum_{\tau} \{g(s_h^{(\tau)}, a_h^{(\tau)}) - f(h_{e_h^{(\tau)}}, \theta)\}^2$. Then, the following inequality holds

$$|g(s_h^{(t)}, a_h^{(t)}) - f(h_{e_h^{(t)}}, \hat{\theta}^g_{h,t})| \le \sigma_g(s_h^{(t)}, a_h^{(t)})$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, with a probability at least $1 - \delta$.

For the purpose of qualifying uncertainty in GLMs, the weighted L_2-norm of ϕ (i.e., $\|\phi_{s,a}\|_{\Lambda_{h,t}^{-1}}$) plays an important role. Because we assume that the Q-function and safety cost function share the same feature, we have a similar lemma on the uncertainty quantifier regarding the Q-function as follows:

Lemma 5.5. Suppose Assumptions 5.2 and 5.3 hold. Let $\hat{\theta}^Q_{h,t} \in \mathbb{R}^d$ denote the ridge estimate; that is, $\hat{\theta}^Q_{h,t} := \arg\min_{\|\theta\|_2 \le 1} \sum_{\tau} \{y_h^{(\tau)} - f(h_{e_h^{(\tau)}}, \theta)\}^2$, where $y_h^{(\tau)} := r(s_h^{(\tau)}, a_h^{(\tau)}) + \max_{a' \in \mathcal{A}} \hat{Q}^{(\tau)}_{r;h+1}(s_{h+1}^{(\tau)}, a')$ for all t with

$$\hat{Q}^{(t)}_{r;h}(s,a) := \min\left\{V_{max}, f(h_{s,a}, \hat{\theta}^Q_{h,t}) + C_{Q-g}\,\sigma_g(s,a)\right\}$$

that is initialized with $\hat{Q}^{(0)}_{r;h} = 0$ for all $h \in [H]$ and $\hat{Q}^{(t)}_{r;H+1} = 0$ for all $1 \le t \le T$. Then, with a universal constant $C_{Q-g} \in \mathbb{R}_{>0}$, the following inequalities holds

$$|Q^*_{r;h}(s_h^{(t)}, a_h^{(t)}) - f(h_{e_h^{(t)}}, \hat{\theta}^Q_{h,t})| \le C_{Q-g}\,\sigma_g(s_h^{(t)}, a_h^{(t)})$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, with a probability at least $1 - \delta$.

Note that $\sigma_g(s_h^{(t)}, a_h^{(t)})$ is the -uncertainty quantifier with respect to the safety cost function. One of the biggest advantages of the generalized linear CMDPs is that the magnitude of uncertainty for the Q-function is proportional to that for the safety cost function. Hence, by exploring the Q-function

based on the optimism in the face of the uncertainty principle [35], the safety cost function is also explored simultaneously, which contributes to the efficient exploration of state-action spaces.

Integration into MASE. The GLM-MASE is an algorithm to integrate the uncertainty quantifiers inferred by the GLM into the MASE sequence. Detailed pseudo code is presented in Appendix E.

To deal with the safety constraint, GLM-MASE leverages the upper bound inferred by the GLM; that is, for all h and t, the agent takes only actions that satisfy

$$f_h(s_h^{(t)}, a_h^{(t)}) + \mathrm{UCB}\big[g_{h,t}(s_h^{(t)}, a_h^{(t)})\big] \le b_h.$$

By Lemma 5.4, such state-action pairs satisfy the safety constraint $g_h(s_h^{(t)}, s_h^{(t)}) \le b_h$, for all h and t, with a probability at least $1-\delta$. If there is no action satisfying the safety constraint (i.e., $A_h^+ = \emptyset$), the emergency stop action a_{stop} is taken, and then the agent receives a penalty defined in (1).

As for policy optimization, we follow the optimism in the face of the uncertainty principle. Specifically, the policy is optimized so that the upper-confidence bound of the Q-function characterized by b is maximized; that is, for any state $s \in \mathcal{S}$, the policy is computed as follows:

$$\pi_h^{(t)}(s) = \arg\max_{a \in \mathcal{A}} Q_{r,h}^{(t)}(s,a).$$

Intuitively, this equation enables us to 1) solve the exploration and exploitation dilemma by incorporating the optimistic estimates of the Q-function and 2) make the agent avoid generating trajectories to violate the safety constraint via the modified reward function.

Theoretical results. We now provide two theorems regarding safety and near-optimality. For both theorems, see Appendix E for the proofs.

Theorem 5.6. Suppose the assumptions in Lemma 5.4 hold. Then, the GLM-MASE satisfies $g(s_h^{(t)}, a_h^{(t)}) \le b_h$ for all $t \in [1, T]$ and $h \in [1, H]$, with a probability at least $1-\delta$.

Theorem 5.7. Suppose the assumptions in Lemmas 5.4 and 5.5 hold. Let C_1 and C_2 be positive, universal constants. Also, with a sufficiently large T, let $t^\star$ denote the smallest integer satisfying $\sigma_{\min}(\Lambda_t) \ge C_1 \sqrt{Hd} + C_2 \sqrt{H \ln \frac{1}{\delta}}$, where $\sigma_{\min}(\Lambda_t)$ is the minimum eigenvalue of the second moment matrix Λ_t. Then, the policy $\pi^{(t)}$ obtained by GLM-MASE at episode t satisfies

$$\sum_{t=t^\star}^{T} \Big[V_r^{\pi^\star} - V_r^{\pi^{(t)}} \Big] \le \widetilde{O}\, H \sqrt{\frac{d^3}{(T - t^\star)}}$$

with probability at least $1-\delta$.

Theorem 5.6 shows that the GLM-MASE guarantees safety with high probability for every time step and episode, which is a variant of Theorem 4.1 under the generalized linear CMDP assumption and corresponding-uncertainty quantifier. Theorem 5.7 demonstrates the agent's ability to act near-optimally after a sufficiently large number of episodes. The proof is based on the following idea. After $t^\star$ episodes, the safety cost function and the Q-function are estimated with an accuracy better than $\frac{\epsilon}{2}$. Then, based on Theorem 4.2, the optimal policy in $\widehat{\mathcal{M}}$ is identical to that in $\mathcal{M}$; thus, the agent achieves a near-optimal policy by leveraging the (well-estimated) optimistic Q-function.

6 A Practical Algorithm

Though we established an algorithm backed by theory under the generalized linear CMDP assumption in Section 5, it is often challenging to obtain proper feature mapping functions in complicated environments. Thus, in this section, we propose a more practical algorithm combining a GP-based estimator to guarantee safety with unconstrained deep RL algorithms to maximize the reward.

Guaranteeing safety via GPs. As shown in the previous sections, the uncertainty quantifier plays a critical role in MASE. To qualify the uncertainty in terms of the safety cost function g, we consider modeling it as a GP, $g(z) \sim \mathcal{GP}(\mu(z), k(z, z'))$, where $z := [s, a]$, $\mu(z)$ is a mean function, and $k(z, z')$ is a covariance function. The posterior distribution over $g(z)$ is computed based on $n \in \mathbb{Z}_{>0}$ observations at state-action pairs $(z_1, z_2, \ldots, z_n)$ with safety measurements $y_n := \{y_1, y_2, \ldots, y_n\}$, where $y_n := g(z_n) + \omega_n$ and $\omega_n \sim \mathcal{N}(0, \sigma^2)$ is zero-mean Gaussian noise

with a standard deviation of $\sqrt{2}R_0$. We consider episodic RL problems, and so $t+h$ for episode t and time step h, although the equality does not hold because of the episode cutoffs. Using the past measurements, the posterior mean, variance, and covariance are computed analytically as $\mu_n(z) = k_n^\top(z)(K_n + \omega^2 I)^{-1} y_n$, $\sigma_n(z) = k_n(z,z)$, and $k_n(z,z') = k(z,z') - k_n^\top(z)(K_n + \omega^2 I)^{-1} k_n(z')$, where $k_n(z) = [k(z_1,z),\dots,k(z_n,z)]^\top$ and K_n is the positive definite kernel matrix. We now present a theorem on the safety guarantee.

Theorem 6.1. Assume $\|g\|_k^2 \le B$ and $N_n \ge 1$ for all $n \ge 1$. Set $\beta_n^{1/2} := B + 4\omega\sqrt{\gamma_n + 1 + \ln(1/\delta)}$ and construct the uncertainty quantifier by

$$\Gamma(s,a) := \beta_n \sigma_n(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{2}$$

where γ_n is the information capacity associated with kernel k. Then, MASE based on (2) satisfies the safety constraint $g(s_h^{(t)}, a_h^{(t)}) \le b_h$ for all t and h with a probability of at least $1-\delta$.

See Appendix F for the proofs. Theorem 6.1 guarantees that the safety constraint is satisfied by combining the GP-based uncertainty quantifier in (2) and the emergency stop action.

Maximizing reward via deep RL. The remaining task is to optimize the policy via the modified reward function $\hat{r}$ in (1), whereby the agent is penalized for emergency stop actions. This problem is decoupled from the safety constraint and can be solved as the following unconstrained RL problem:

$$\pi := \arg\max_\pi V_{\hat{r}}^\pi. \tag{3}$$

There are many excellent algorithms for solving (3), such as trust region policy optimization (TRPO, [31]) and twin delayed deep deterministic policy gradient (TD3, [17]). One of the key benefits of our MASE is such compatibility with a broad range of unconstrained (deep) RL algorithms.

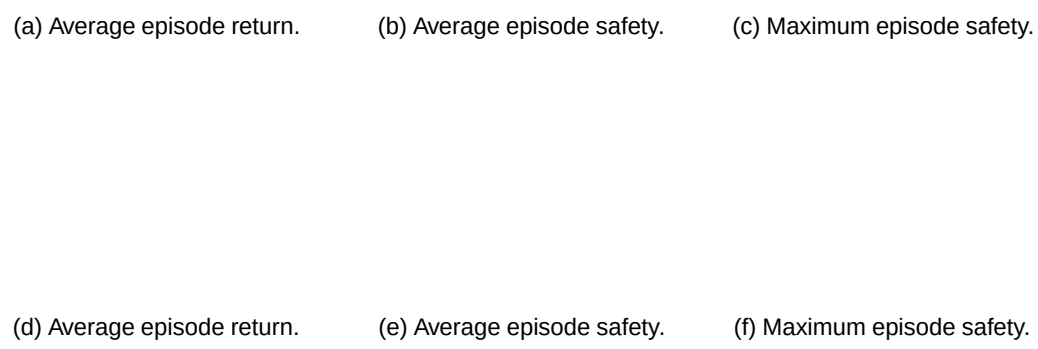
7 Experiments

We conduct two experiments. The first is on Safety Gym [26], where an agent must maximize the expected cumulative reward under a safety constraint with additive structures as in Problems 1 and 2. The safety cost function g is binary (i.e., 1 for an unsafe state-action pair and 0 otherwise), and the safety threshold is set to $b_1 = 20$. The reason for choosing Safety Gym is that this benchmark is complex and elaborate, and has been used to evaluate a variety of excellent algorithms. The second is a grid world where a safety constraint is instantaneous as in Problem 3. Due to the page limit, we present the settings and results of the grid-world experiment in Appendix H.

To solve the Safety Gym tasks, we implement the practical algorithm presented in Section 6 as follows. First, we convert the problem into a MASE problem by defining $b_h := \gamma_g^{-1}\left(\frac{1}{h}\sum_{h'=0}^{h} \gamma_g^{h'} g(s_{h'}, a_{h'})\right)$ and enforcing the safety constraint represented as $g(s_h, a_h) \le b_h$ for every time step h in each episode. Second, to infer the safety cost, we use deep GP to conduct training and inference, as in [30], which is known for dealing with high-dimensional input spaces. Finally, as for the policy optimization in (3), we leverage the TRPO algorithm. We delegate other details to Appendix G.

Baselines and metrics. We use the following four algorithms as baselines. The first is TRPO, which is a safety-agnostic deep RL algorithm that purely optimizes a policy without safety consideration. The second and third are CPO [1] and TRPO-Lagrangian [28], which are well-known algorithms for solving CMDPs. The final algorithm is Sauté RL [32], which is a recent, state-of-the-art algorithm for solving safe RL problems where constraints must be satisfied almost surely. We employ the following three metrics to evaluate our MASE and the aforementioned four baselines: 1) the expected cumulative reward, 2) the expected cumulative safety, and 3) the maximum cumulative safety. We execute each algorithm with five random seeds and compute the means and confidence intervals.

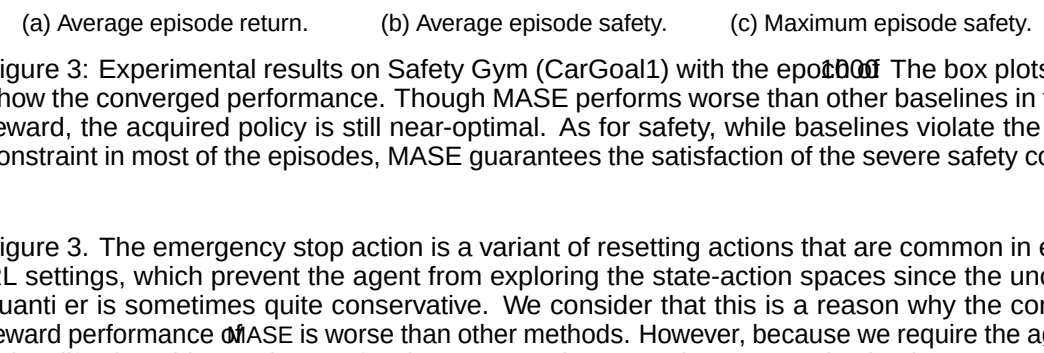
Results. The experimental results are summarized in Figure 2. The figures show that TRPO, TRPO-Lagrangian, and CPO successfully learn the policies, but violate the safety constraints during training and even after convergence. Sauté RL is much safer than those three algorithms, but the safety constraint is not satisfied in some episodes, and the performance of the policy significantly deteriorates in terms of the cumulative reward during training. MASE obtains better policies in a smaller number of samples compared with Sauté RL, while also satisfying the safety constraints with respect to both the average and the worst-case. Note that, after convergence, the policy obtained by MASE performs worse than those obtained by the baseline algorithms in terms of reward, as shown in



(a) Average episode return. (b) Average episode safety. (c) Maximum episode safety.

(d) Average episode return. (e) Average episode safety. (f) Maximum episode safety.

Figure 2: Experimental results on Safety Gym (Top: PointGoal1, Bottom: CarGoal1). The proposed MASE satisfies the safety constraint in every episode and achieves better performance in terms of the reward than the state-of-the-art method called Sauté RL. Conventional methods (i.e., TRPO, TRPO-Lagrangian, and CPO) repeatedly violate the safety constraint, especially in the early phase of training. Shaded areas represent 95% confidence intervals across five different random seeds.



(a) Average episode return. (b) Average episode safety. (c) Maximum episode safety.

Figure 3: Experimental results on Safety Gym (CarGoal1) with the epoch of 1000. The box plots show the converged performance. Though MASE performs worse than other baselines in terms of reward, the acquired policy is still near-optimal. As for safety, while baselines violate the safety constraint in most of the episodes, MASE guarantees the satisfaction of the severe safety constraint.

Figure 3. The emergency stop action is a variant of resetting actions that are common in episodic RL settings, which prevent the agent from exploring the state-action spaces since the uncertainty quantifier is sometimes quite conservative. We consider that this is a reason why the converged reward performance of MASE is worse than other methods. However, because we require the agent to solve difficult problems where safety is guaranteed at every time step and episode, we consider that this result is reasonable, and further performance improvements are left to future work.

8 Conclusion

In this article, we first introduced the GSE problem and proved that it is more general than three common safe RL problems. We then proposed MASE to optimize a policy under safety constraints that allow the agent to execute an emergency stop action at the sacrifice of a penalty based on the λ-uncertainty quantifier. As a specific instance of MASE, we first presented GLM-MASE to theoretically guarantee the near-optimality and safety of the acquired policy under generalized linear CMDP assumptions. Finally, we provided a practical MASE and empirically evaluated its performance in comparison with several baselines on the Safety Gym and grid-world.

Acknowledgments and Disclosure of Funding

We would like to thank the anonymous reviewers for their helpful comments. This work is partially supported by JST CREST JPMJCR201 and by JSPS KAKENHI Grant 21K14184.

References

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

[2] Alshiekh, M., Bloem, R., Ehlers, R., Könighofer, B., Niekum, S., and Topcu, U. (2018). Safe reinforcement learning via shielding. In AAAI Conference on Artificial Intelligence (AAAI).

[3] Altman, E. (1999). Constrained Markov decision processes, volume 7. CRC Press.

[4] Amani, S., Alizadeh, M., and Thrampoulidis, C. (2019). Linear stochastic bandits under safety constraints. In Neural Information Processing Systems (NeurIPS).

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

[6] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

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

[8] Bajracharya, M., Maimone, M. W., and Helmick, D. (2008). Autonomy for mars rovers: Past, present, and future. Computer, 41(12):44–50.

[9] Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., and Schoellig, A. P. (2022). Safe learning in robotics: From learning-based control to safe reinforcement learning. Annual Review of Control, Robotics, and Autonomous Systems, 5:411–444.

[10] Carr, S., Jansen, N., Junges, S., and Topcu, U. (2023). Safe reinforcement learning via shielding under partial observability. In AAAI Conference on Artificial Intelligence.

[11] 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.

[12] Chowdhury, S. R. and Gopalan, A. (2017). On kernelized multi-armed bandits. In International Conference on Machine Learning (ICML).

[13] Ding, D., Zhang, K., Basar, T., and Jovanovic, M. (2020). Natural policy gradient primal-dual method for constrained Markov decision processes. In Neural Information Processing Systems (NeurIPS).

[14] 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.

[15] Eysenbach, B., Gu, S., Ibarz, J., and Levine, S. (2018). Leave no trace: Learning to reset for safe and autonomous reinforcement learning. In International Conference on Learning Representations (ICLR).

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

[17] Fujimoto, S., van Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods. In International Conference on Machine Learning (ICML).

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

[19] Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably ef cient reinforcement learning with linear function approximation. In Conference on Learning Theory (COLT).

[20] Könighofer, B., Bloem, R., Junges, S., Jansen, N., and Serban, A. (2020). Safe reinforcement learning using probabilistic shields. In International Conference on Concurrency Theory: CONCUR.

[21] Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In International Conference on World Wide Web (WWW).

[22] Li, L., Lu, Y., and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In International Conference on Machine Learning (ICML), pages 2071–2080.

[23] Melcer, D., Amato, C., and Tripakis, S. (2022). Shield decentralization for safe multi-agent reinforcement learning. In Neural Information Processing Systems (NeurIPS).

[24] Ono, M., Pavone, M., Kuwata, Y., and Balaram, J. (2015). Chance-constrained dynamic programming with application to risk-aware robotic space exploration. Autonomous Robots, 39(4):555–571.

[25] Paternain, S., Chamon, L., Calvo-Fullana, M., and Ribeiro, A. (2019). Constrained reinforcement learning has zero duality gap. In Neural Information Processing Systems (NeurIPS).

[26] Pfrommer, S., Gautam, T., Zhou, A., and Sojoudi, S. (2022). Safe reinforcement learning with chance-constrained model predictive control. In Learning for Dynamics and Control Conference (L4DC), pages 291–303.

[27] Rasmussen, C. E. (2003). Gaussian processes in machine learning. In Summer school on machine learning, pages 63–71. Springer.

[28] Ray, A., Achiam, J., and Amodei, D. (2019). Benchmarking safe exploration in deep reinforcement learning. OpenAI.

[29] Rockafellar, R. T., Uryasev, S., et al. (2000). Optimization of conditional value-at-risk. Journal of risk, 2:21–42.

[30] Salimbeni, H. and Deisenroth, M. (2017). Doubly stochastic variational inference for deep Gaussian processes. In Neural Information Processing Systems (NeurIPS).

[31] Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In International Conference on Machine Learning (ICML).

[32] Sootla, A., Cowen-Rivers, A. I., Jafferjee, T., Wang, Z., Mguni, D. H., Wang, J., and Ammar, H. (2022). Sauté RL: Almost surely safe reinforcement learning using state augmentation. In International Conference on Machine Learning (ICML).

[33] Stooke, A., Achiam, J., and Abbeel, P. (2020). Responsive safety in reinforcement learning by pid Lagrangian methods. In International Conference on Machine Learning (ICML), pages 9133–9143.

[34] Strehl, A. L. and Littman, M. L. (2005). A theoretical analysis of model-based interval estimation. In International Conference on Machine Learning (ICML).

[35] 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.

[36] Sui, Y., Gotovos, A., Burdick, J. W., and Krause, A. (2015). Safe exploration for optimization with Gaussian processes. In International Conference on Machine Learning (ICML).

[37] Sun, H., Xu, Z., Fang, M., Peng, Z., Guo, J., Dai, B., and Zhou, B. (2021). Safe exploration by solving early terminated MDP. arXiv preprint arXiv:2107.04200.

[38] Thananjeyan, B., Balakrishna, A., Nair, S., Luo, M., Srinivasan, K., Hwang, M., Gonzalez, J. E., Ibarz, J., Finn, C., and Goldberg, K. (2021). Recovery RL: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 6(3):4915–4922.

[39] Thomas, G., Luo, Y., and Ma, T. (2021). Safe reinforcement learning by imagining the near future. In *Neural Information Processing Systems (NeurIPS)*

[40] Turchetta, M., Kolobov, A., Shah, S., Krause, A., and Agarwal, A. (2020). Safe reinforcement learning via curriculum induction. In *Neural Information Processing Systems (NeurIPS)* volume 33, pages 12151–12162.

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

[42] Wachi, A., Sui, Y., Yue, Y., and Ono, M. (2018). Safe exploration and optimization of constrained MDPs using Gaussian processes. In *AAAI Conference on Artificial Intelligence (AAAI)*

[43] Wachi, A., Wei, Y., and Sui, Y. (2021). Safe policy optimization with local generalized linear function approximations. In *Neural Information Processing Systems (NeurIPS)*

[44] Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. (2020). Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations (ICLR)*

[45] Wang, Y., Zhan, S. S., Jiao, R., Wang, Z., Jin, W., Yang, Z., Wang, Z., Huang, C., and Zhu, Q. (2023). Enforcing hard constraints with soft barriers: Safe reinforcement learning in unknown stochastic environments. In *International Conference on Machine Learning*, pages 36593–36604. PMLR.

[46] Yang, L. and Wang, M. (2019). Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning (ICML)*

Appendices

A Limitations and Potential Negative Societal Impacts

We will first discuss limitations and potential negative societal impacts regarding our work.

A.1 Limitations

One of the major limitations of this study is that emergency stop actions are allowed for agents. Emergency stop actions should often be avoided because of the expensive cost of human intervention in many applications (e.g., the agent is in a hazardous or remote environment). In future work, we will investigate an algorithm that requires the agent to learn a reset policy allowing them to return to the initial state as in [15], rather than asking for human intervention via emergency stop actions.

Another limitation is how to construct the uncertainty quantifier. In our experiment, because we used a computationally inexpensive deep GP algorithm [30] and the uncertainty quantifier is updated at the end of the episode (see Line 13 in Algorithm 1), the computational time of the GP part was much smaller than the RL part in our experiment settings. However, since GP is generally a computationally expensive algorithm, GP can be a computational bottleneck in some cases.

A.2 Potential Negative Societal Impacts

We believe that safety is an essential requirement for applying RL in many real problems. While we have not found any potential negative societal impact of our proposed method, we must remain aware that RL algorithms are vulnerable to misuse and ours is no exception.

B Proof of Theorem 3.1

We first present lemmas regarding the relationship between the GSE problem and Problems 1, 2, and 3. After that, we present the proof for the Theorem 3.1 in Appendix B.4 by combining them.

B.1 Relationship between the GSE problem and Problem 1

Lemma B.1. Problem 1 can be transformed into the GSE problem.

Proof. We first utilize a safety state augmentation technique presented in Sootla et al. [32] by defining a new variable ξ_h such that

$$\xi_h := \gamma_g^h \left(1 - \sum_{h'=1}^{h} \gamma_g^{-h'} g(s_{h'}, a_{h'}) \right), \quad \forall h \in [1, H]. \tag{4}$$

This new variable ξ_h means the remaining safety budget associated with the discount factor γ_g, which is updated as follows:

$$\xi_{h+1} = \gamma_g^{-1} \left(\xi_h - g(s_h, a_h) \right) \quad \text{with} \quad \xi_0 = \xi_1. \tag{5}$$

By (4), the necessary and sufficient condition for satisfying the constraint in Problem 1 is

$$\xi_h \geq 0, \quad \forall h \in [1, H]. \tag{6}$$

By (5), we have

$$\xi_{h+1} \geq 0, \quad \forall h \in [1, H] \iff \xi_h - g(s_h, a_h) \geq 0, \quad \forall h \in [1, H]. \tag{7}$$

In summary, by introducing the new variable ξ_h, Problem 1 is rewritten to the following problem:

$$\max V_r \quad \text{subject to} \quad \Pr[\xi_h - g(s_h, a_h) \geq 0 \mid \xi_h, \pi] = 1, \quad \forall h \in [1, H]. \tag{8}$$

The aforementioned problem (8) is a special case of the GSE problem with $b_h := \xi_h$. Therefore, we obtain the desired lemma. □

B.2 Relationship between the GSE problem and Problem 2

Lemma B.2. Problem 2 can be transformed into the GSE problem.

Proof. The following chain of inequalities holds:

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{h=1}^{H} {}_{g}\mathbb{I}(s_h \in S_{\text{unsafe}}) \mid P; \quad {}_{2}\right] \\
\Leftrightarrow\ & \mathbb{E}\left[\sum_{h'=1}^{h} {}_{g}\mathbb{I}(s_{h'} \in S_{\text{unsafe}}) \mid P; \quad {}_{2}; \quad \forall h \in [1; H]\right] \\
\Leftrightarrow\ & \mathbb{E}\left[{}_{g}\mathbb{I}(s_h \in S_{\text{unsafe}}) \mid P; \quad {}_{2} \quad \mathbb{E}\left[\sum_{h'=1}^{h-1} {}_{g}\mathbb{I}(s_{h'} \in S_{\text{unsafe}}) \mid P; \quad ; \quad \forall h \in [2; H]\right]\right] \\
\Leftrightarrow\ & \mathbb{E}[\mathbb{I}(s_h \in S_{\text{unsafe}}) \mid P; \quad] \quad {}_{g}{}^{h} \quad {}_{2} \quad \mathbb{E}\left[\sum_{h'=1}^{h-1} {}_{g}\mathbb{I}(s_{h'} \in S_{\text{unsafe}}) \mid P; \quad ; \quad \forall h \in [2; H]\right]\right)
\end{aligned}
$$

Because we assume Markov property, we simply define the safety cost function $g : S \times A \to \mathbb{R}$ as

$$g(s_h; a) := \mathbb{E}[\mathbb{I}(s_h \in S_{\text{unsafe}}) \mid P; \quad]; \quad \forall a \in A.$$

We now set

$$b_h := {}_{g}{}^{h} \quad {}_{2} \quad \mathbb{E}\left[\sum_{h'=1}^{h-1} {}_{g}\mathbb{I}(s_{h'} \in S_{\text{unsafe}}) \mid P; \quad ;\right)$$

then the Problem 2 can be transformed into

$$\Pr[\, g(s; a) \quad b_h \mid P; \quad] = 1.$$

Finally, we obtained the desired lemma. □

B.3 Relationship between the GSE problem and Problem 3

Lemma B.3. Problem 3 can be transformed into the GSE problem.

Proof. Set $b_h = {}_{3}$ for all h. Then, the GSE problem is identical to Problem 3. □

B.4 Summary

Proof. Combining Lemma B.1, B.2, and B.3, we obtain the desired Theorem 3.1. □

C Connections to Safe RL Problems with Chance Constraints

As a strongly related formulation to Problem 2, policy optimization under joint chance-constraints has been studied especially in the field of control theory such as Ono [24] and Pfrommer et al. [26], which is written as

Problem 5 (Safe RL with joint chance constraints). Let ${}_{5} \in \mathbb{R}_{\geq 0}$ be a constant representing a safety threshold. Also, let $S_{\text{unsafe}} \subseteq S$ denote a set of unsafe states. Find the optimal policy π such that

$$\max \ V_r \quad \text{subject to} \quad \Pr\left[\bigvee_{h=1}^{H} s_h \in S_{\text{unsafe}} \mid P; \quad \right] \quad {}_{5}.$$

Lemma C.1. Problem 2 is a conservative approximation of Problem 5.

Proof. This lemma mostly follows from Theorem 1 in Ono et al. [24]. Regarding the constraint in Problem 5, we have the following chain of equations:

$$\Pr\left[\bigcap_{h=1}^{H} s_h \in S_{\text{unsafe}} \mid P, \right] \le \sum_{h=1}^{H} \Pr[s_h \in S_{\text{unsafe}} \mid P,] \tag{9}$$

$$= \sum_{h=1}^{H} \mathbb{E}[I(s_h \in S_{\text{unsafe}}) \mid P,] = \mathbb{E}\left[\sum_{h=1}^{H} I(s_h \in S_{\text{unsafe}}) \mid P, \right]. \tag{10}$$

In the first step, we used Boole's inequality (i.e., $\Pr[A \cap B] \le \Pr[A] + \Pr[B]$). The final term in (10) is the LHS of the constraint in Problem 2, which implies that Problem 2 is a conservative approximation of Problem 5. Therefore, the GSE problem is also a conservative approximation of Problem 5. □

Corollary C.2. Suppose we solve the GSE problem by properly defining the safety function $g(\cdot, \cdot)$ and the threshold b_h. Then, the obtained policy is a conservative solution of Problem 5.

Remark C.3. It is extremely challenging to directly solve the Problem 5 characterized by joint chance-constraints without approximation, as discussed in Ono et al. [24]. Practically, it would be a promising and realistic approach to solve Problem 5 by converting it into the GSE problem.

More detailed explanations for the aforementioned remark are as follows. It is extremely challenging to directly solve the Problem 5 characterized by joint chance-constraints. Most of the previous work does not directly deal with this type of constraint and uses some approximations or assumptions. For example, Pfrommer et al. [26] assume a known linear time-invariant dynamics. Also, Ono et al. [24] approximate the joint chance constraint as in the above procedure and obtain

$$\Pr\left[\bigcap_{h=1}^{H} s_h \in S_{\text{unsafe}} \mid P, \right] \le \mathbb{E}\left[\sum_{h=1}^{H} I(s_h \in S_{\text{unsafe}}) \mid P, \right].$$

This is a conservative approximation with an additive structure, which is easier to solve than the original joint chance constraint. Ono et al. [24] deals with the above constraints with additive safety structure. By additionally transforming the conservatively-approximated problem into the GSE problem, the problem would become easier to handle because the safety constraint is instantaneous.

D Proof of Theorems 4.1 and 4.2

D.1 Proof of Theorem 4.1

Proof. Recall that, at every time step h, the MASE chooses actions satisfying

$$(s_h, a_h) + (s_h, a_h) \le b_h. \tag{11}$$

By definition of the $$-uncertainty quantifier, the actions that are conservatively chosen based on (11) also satisfy the safety constraint, with a probability of at least 1; that is,

$$g(s_h, a_h) \le b_h. \tag{12}$$

In addition, when there is no action satisfying (11), the emergency stop action is executed; that is, safety is guaranteed with a probability 1. Hence, the desired theorem is now obtained. □

D.2 Proof of Theorem 4.2

Proof. Assumption 3.2 implies that there exists a policy that satisfies a more conservative safety constraint written as

$$g(s_h, a_h) \le b_h - , \quad \forall h \in [1, H]. \tag{13}$$

By combining (13) and the assumption $(s, a) \le \frac{1}{2}$, $\forall (s, a) \in S \times A$, we have

$$(s_h, a_h) + (s_h, a_h) \le g(s_h, a_h) + b_h, \quad \forall h \in [1, H], \tag{14}$$

Algorithm 2 GLM-MASE

for episode $t = 1, 2, \ldots, T$ do
  for time $h = 1, \ldots, H$ do
    Take action $a_h = \pi_h(s_h)$ within A_h^+
    Receive reward $r(s_h, a_h)$ and next state s_{h+1}
    Receive safety cost $g(s_h, a_h)$ and update safety threshold b_{h+1}
    if $A_h^+ = \emptyset$ then
      Compute $b(s_h, a_h) = \frac{c}{\min_{a \in A} \hat{g}(s_{h+1}, a)}$
      Append $(s_h, a_h, b(s_h, a_h), s_{h+1})$ to D
      break (i.e., emergency stop action)
    else
      Append $(s_h, a_h, r(s_h, a_h), s_{h+1})$ to D
  Update $b_{h,t}^g$ and $b_{h,t}^Q$
  Compute the optimistic Q-estimate

$$Q_{r,h}^{(t)}(s, a) = \min\{V_{max}, f(\hat{r}_{s,a}, \hat{Q}_{h,t}) + C_{Q=g}\,\epsilon(s, a)\}$$

  Optimize the policy by

$$\pi_h^{(t)}(s) = \arg\max_{a \in A} Q_{r,h}^{(t)}(s, a).$$

  Update $\epsilon(s, a) := C_g \|\phi_{s,a}\|_{\Gamma_{h,t}^{-1}}$ and then rewrite D

which guarantees that there exists a policy that conservatively satisfies the safety constraint via the ϵ-uncertainty quantifier $\epsilon(\cdot, \cdot)$ at every time step h, with a probability of at least $1 - \delta$.

When we set c to be a sufficiently large scalar such that $c > \frac{1}{2 - \frac{H}{r}} V_{max}$, the penalty b satisfies

$$\begin{aligned} b(s_h, a_h) &= \frac{c}{\min_{a \in A} \hat{g}(s_{h+1}, a)} \\ &< \frac{\frac{1}{2} - \frac{1}{\frac{H}{r}} V_{max}}{\frac{1}{2}} \\ &< {}_r^H V_{max}. \end{aligned}$$

This means that, when the constraint violation happens even a single time, the value by a policy obtained in M' becomes negative because $\max_s V_r(s) - V_{max}$.

Under Assumption 3.2, after convergence, the optimal policy in M' will not violate the safety constraint, and thus the emergency stop action will not be executed. In this case, the modified (unconstrained) MDP M' is identical to the original CMDP M. Therefore, we now obtain the desired theorem. □

E Supplementary materials regarding GLM-MASE

E.1 Pseudo-code for GLM-MASE

We first present the pseudo-code for GLM-MASE in Algorithm 2.

E.2 Proofs of Lemmas 5.4 and 5.5

Proof. See Lemma 1 (and Lemma 7) in Wang et al. [44]. □

E.3 Preliminary Lemmas

Lemma E.1. Suppose the assumptions in Lemma 5.4 and 5.5 hold. Let C_1 and C_2 be positive, universal constants. Also, with a sufficiently large T, let t^* denote the smallest integer satisfying

$$\epsilon_{min}(t) \le C_1 \sqrt{\frac{Hd}{t}} + C_2 \sqrt{\frac{H \ln \frac{1}{\delta}}{t}} \le \frac{1}{2} C_g \qquad (15)$$

where $\lambda_{\min}(\cdot)$ *is the minimum eigenvalue of the second moment matrix* $\cdot$ *. Then, we have*

$$(s_h^{(t)}, a_h^{(t)}) \quad \frac{1}{2} \quad : \tag{16}$$

Proof. By Proposition 1 of Li et al. [22],

$$\lambda_{\min}(\quad_{h,t}) \quad \lambda_{\min}(\quad)tH \quad C_1\sqrt{tHd} \quad C_2\sqrt{tH\ln\quad^{1}}.$$

By combining the aforementioned inequality with (15), we have

$$\lambda_{\min}(\quad_{h,t}) \quad 2 \quad C_g \quad \quad ^{1}.$$

Using the definition of $\lambda_{\max}(\quad_{h,t}^{1}) = \frac{1}{\lambda_{\min}(\quad_{h,t})}$, the following chain of equations hold for all $t \in [t^{?}, T]$ and $h \in [1, H]$:

$$\begin{aligned}(s_h^{(t)}, a_h^{(t)}) &= C_g \quad k \quad_{s,a}k_{\quad_{h,t}^{1}} \\ &\quad C_g \quad \lambda_{\max}(\quad_{h,t}^{1}) \\ &= C_g \quad \lambda_{\min}^{1}(\quad_{h,t}) \\ &\quad \frac{1}{2} \quad .\end{aligned}$$

Therefore, we have the desired lemma. □

E.4 Proof of Theorem 5.6

Proof. By definition, the GLM-MASE chooses actions satisfying

$$f(h\ e_h^{(\ )}, b_{h,t}^{g}\ i) + \quad (s_h^{(t)}, a_h^{(t)}) \quad b_h. \tag{17}$$

By Lemma 5.4, the actions that are conservatively chosen based on (17) also satisfy the actual safety constraint, with a probability at least $1 \quad$; that is,

$$g(s_h^{(t)}, a_h^{(t)}) \quad b_h. \tag{18}$$

In addition, when there is no action satisfying (17), the emergency stop action is executed where no unsafe action will be executed. Hence, the desired theorem is now obtained. □

E.5 Proof of Theorem 5.7

By Assumption 3.2, the optimal policy $^{?}$ satisfies

$$g(s_h, \ ^{?}(s_h)) \quad b_h \quad ; \quad 8h \in [1, H]. \tag{19}$$

Thus, the set of state-action pairs that are potentially visited by $^{?}$ are written as

$$f(s,a) \in S \quad A \mid g(s,a) \quad b_h \quad g,$$

which satisfies the following chain of inequalities:

$$\begin{aligned}f(s,a) \mid g(s,a) \quad b_h \quad g \quad & f(s,a) \mid \quad (s,a) \quad (s,a) \quad b_h \quad g \\ & f(s,a) \mid \quad (s,a) + \quad (s,a) \quad b_h\, g.\end{aligned}$$

The state and action spaces in the last line represent the set of state-action pairs that may be visited by the policy obtained by the MASE algorithm. We used Lemma 5.4 in the first line and $(s,a) < \frac{}{2}$ in the second line.

By Lemma 8 in Strehl and Littman [34], the total regret can be decomposed into two parts as follows:

$$\sum_{t=t^{?}}^{\bar{} \quad} \left[V_{r\ \ ^{?}} \quad V_{r\ \ ^{t}} \right] = R(T) + \sum_{t=t^{?}}^{} V_{\max}\ ; \tag{20}$$

where the first term is the regret under the condition that the confidence bound based on -uncertainty quantifier successfully contains the true safety function. Also, the second term is the regret under the opposite condition, which occurs with a probability .

The first term in (20) is upper-bounded based on Wang et al. [44] as follows:

$$R(T) \quad O \quad H \sqrt{(T \quad t^?)\ln((T \quad t^?)H)} + H^{-} \sqrt{M + - + d^2 \ln \frac{- + \max}{(T \quad t^?)H} \quad (T \quad t^?)d \ln 1 + \frac{(T \quad t^?)}{d}} \quad \tilde{O}(H \sqrt{d^3(T \quad t^?)}):$$

As for the second term in (20), set $= \frac{1}{TH}$ and then we have

$$\begin{aligned} \sum_{t=t^?} V_{\max} &= \sum_{t=t^?} V_{\max} \frac{1}{TH} \\ &= \sum_{t=t^?} \frac{1}{1} \frac{H}{} \frac{1}{TH} \\ &\sum_{t=t^?} H \frac{1}{TH} \\ &O(1): \end{aligned}$$

In summary, the regret can be upper bounded by $\tilde{O}(H \sqrt{d^3(T \quad t^?)})$. We now obtain the desired theorem.

F Proof of Theorem 6.1

Lemma F.1 (-uncertainty quantifier)**.** *Assume* $kgk_k^2 \quad B$ *and* $N_n \quad !$ *for all* $n \quad 1$*. Set*

$$(s;a) := \quad _n \quad _n(s;a); \quad 8(s;a) \; 2 \; S \quad A \tag{21}$$

with $_n^{1=2} := B + 4! \sqrt{_n + 1 + \ln \quad ^1}$*, where* $_n$ *is the information capacity associated with kernel* k*. Then, is a -uncertainty quantifier.*

Proof. Recall the assumption that $kgk_k^2 \quad B$ and $N_n \quad !, 8n \quad 1$. Also, set

$$_n^{1=2} := B + 4! \sqrt{_n + 1 + \ln \quad ^1}:$$

By Theorem 2 in Chowdhury and Gopalan [12], we have

$$j g(s;a) \quad _n(s;a) j \quad _n \quad _n(s;a); \quad 8(s;a) \; 2 \; S \quad A$$

for all $n \quad 1$, with probability at least $1 \quad$. Now, define

$$(s;a) := \quad _n \quad _n(s;a); \quad 8(s;a) \; 2 \; S \quad A \tag{22}$$

then we interpret that $: S \quad A \; ! \; \mathbb{R}$ is a -uncertainty quantifier based on GP. □

F.1 Proof of Theorem 6.1

Proof. Based on Lemma F.1, when there is at least one safe action (i.e., $A_h^+ \neq ;$), the satisfaction of the safety constraint is guaranteed based on the -uncertainty quantifier with a probability at least $1 \quad$. Also, if there is no safe action (i.e., $A_h^+ = ;$), the emergency stop action is executed. In both cases, MASE guarantees the satisfaction of the safety constraint with probability at least $1 \quad$. □

Table 1: Hyper-parameters for Safety Gym experiments.

	NAME	VALUE
COMMON PARAMETERS	NETWORK ARCHITECTURE	[64; 64]
	ACTIVATION FUNCTION	tanh
	LEARNING RATE (CRITIC)	$5 \quad 10^{-3}$
	LEARNING RATE (POLICY)	$3 \quad 10^{-4}$
	LEARNING RATE (PENALTY)	$3 \quad 10^{-2}$
	DISCOUNT FACTOR (REWARD)	0:99
	DISCOUNT FACTOR (SAFETY)	0:99
	STEPS PER EPOCH	10; 000
	NUMBER OF GRADIENT STEPS	80
	NUMBER OF EPOCHS	500
	TARGET KL	0:01
TRPO & CPO	DAMPING COEFFICIENT	0:1
	BACKTRACK COEFFICIENT	0:8
	BACKTRACK ITERATIONS	10
	LEARNING MARGIN	FALSE
MASE	PENALTY FOR EMERGENCY STOP ACTIONS	1
	DEEP GP NETWORK ARCHITECTURE	[16; 16]
	NUMBER OF INDUCING POINTS	128
	KERNEL FUNCTION	RADIAL BASIS FUNCTION

G Details of Safety-Gym Experiment

We present the details regarding our experiments using Safety-Gym. Our experimental setting is based on Sootla et al. [32], which is slightly different from the original Safety Gym in that the obstacles (i.e., unsafe region) are replaced deterministically so that the environment is solvable and there is a viable solution. In this experiment, we used a machine with Intel(R) Xeon(R) Silver 4210 CPU, 128GB RAM, and NVIDIA A100 GPU. For a fair comparison, we basically used the same hyper-parameter as in Sootla et al. [32], which is summarized in Table 1.

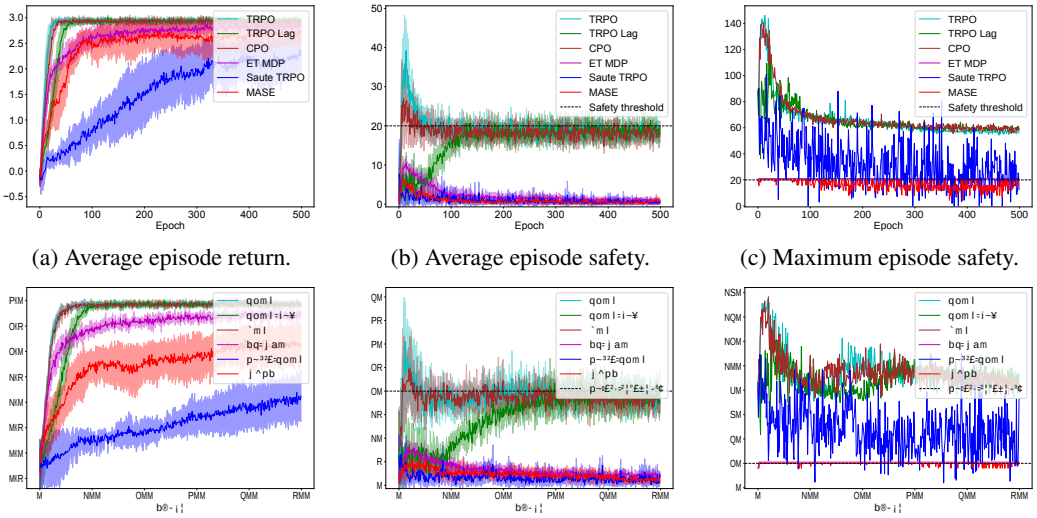
In our experiment, when the agent identified that there was no safe action based on the GP-based uncertainty quantifier, we simply terminated the current episode (i.e., resetting) immediately after the emergency stop action and started the new episode. The frequency of the emergency stop actions is shown in Table 2.

Table 2: Frequency of the emergency stop actions.

TASK	TOTAL	LAST 100 EPISODES
POINTGOAL1	154/500	24/100
CARGOAL1	397/500	46/100

The emergency stop action is a variant of so-called resetting actions that are common in episodic RL settings, which prevent the agent from exploring the state-action spaces since the uncertainty quantifier is sometimes quite conservative. We consider that this is the reason why the reward performance of our MASE is worse than other methods (e.g., TRPO-Lagrangian, CPO). However, because we require the agent to solve more difficult problems where safety is guaranteed at every time step and episode, we consider that this result is reasonable to some extent. Though it is better for an algorithm for such a severe safety constraint to have a comparable performance as CPO, we will leave it for future work.

We also conducted an experiment to compare the performance of MASE with the early-terminated MDP (ET-MDP, [37]) algorithm. The ET-MDP is an algorithm to execute emergency stop actions *immediately after* safety constraints are violated. Figure 4 shows the experimental results. The ET-MDP and MASE exhibit similar learning curves on the average episode reward and average episode safety. However, while ET-MDP violated the safety constraint in most episodes (i.e., almost



(a) Average episode return. (b) Average episode safety. (c) Maximum episode safety.

(d) Average episode return. (e) Average episode safety. (f) Maximum episode safety.

Figure 4: Experimental results on Safety Gym (Top: PointGoal1, Bottom: CarGoal1) with an additional implementation of the Early-terminated MDP (ET-MDP) algorithm (Sun et al. [37]).

all episodes are terminated after an unsafe action is executed), MASE did not violate any safety constraint.

H Grid-world Experiment

We also conduct an experiment using the grid-world problem as in Wachi and Sui [41]. Experimental settings are based on their original implementation (`https://github.com/akifumi-wachi-4/safe_near_optimal_mdp`). We consider a 20 20 square grid in which reward and safety functions are randomly generated. There are two types regarding the safety threshold: one is time-invariant as in [41] and the other is time-variant as in the GSE problem.

We run SNO-MDP [41] and MASE in 100 randomly generated environments, and we compute the reward collected by the algorithms and count the number of episodes in which the safety constraint is violated at least once. The reward is normalized with respect to that by SNO-MDP.

Table 3: Experimental results for grid-world experiments.

	TIME-INVARIANT SAFETY THRESHOLD		TIME-VARIANT SAFETY THRESHOLD	
	REWARD	SAFETY VIOLATION	REWARD	SAFETY VIOLATION
SNO-MDP [41]	1:0 0:0	0	1:0 0:0	87
MASE (OURS)	1:0 0:0	0	2:4 1:0	0

The experimental results are shown in Table 3. When the safety threshold is time-invariant, MASE behaves identically with SNO-MDP; thus, the performance of the MASE is comparable with that of SNO-MDP. When the safety threshold is time-variant, SNO-MDP cannot deal with it by nature; hence, the safety constraint is not satisfied in most of the episodes. In contrast, our MASE satisfies the safety constraint in every episode, which also contributes to the larger reward.