Joint Synthesis of Safety Certificate and Safe Control Policy using Constrained Reinforcement Learning

Haitong Ma   MAHT19@MAILS.TSINGHUA.EDU.CN
School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China

Changliu Liu   CLIU6@ANDREW.CMU.EDU
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Shengbo Eben Li   LISHBO@TSINGHUA.EDU.CN

Sifa Zheng   ZSF@TSINGHUA.EDU.CN
School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China

Jianyu Chen   JIANYUCHEN@TSINGHUA.EDU.CN
Institute of Interdisciplinary Information Science, Tsinghua University, Beijing 100084, China
Shanghai Qizhi Institute, Shanghai 200000, China

Editors: R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, M. Kochenderfer

Abstract

Safety is the major consideration in controlling complex dynamical systems using reinforcement learning (RL), where the *safety certificates* can provide provable safety guarantees. A valid safety certificate is an energy function indicating that safe states are with low energy, and there exists a corresponding *safe control policy* that allows the energy function to always dissipate. The safety certificates and the safe control policies are closely related to each other and both challenging to synthesize. Therefore, existing learning-based studies treat either of them as prior knowledge to learn the other, limiting their applicability to general systems with unknown dynamics. This paper proposes a novel approach that simultaneously synthesizes the energy-function-based safety certificates and learns the safe control policies with constrained reinforcement learning (CRL). We do not rely on prior knowledge about *either* a prior control law *or* a perfect safety certificate. In particular, we formulate a loss function to optimize the safety certificate parameters by minimizing the occurrence of energy increases. By adding this optimization procedure as an outer loop to the Lagrangian-based CRL, we jointly update the policy and safety certificate parameters, and prove that they will converge to their respective local optima, the optimal safe policies and valid safety certificates. Finally, we evaluate our algorithms on multiple safety-critical benchmark environments. The results show that the proposed algorithm learns solidly safe policies with no constraint violation. The validity or feasibility of synthesized safety certificates is also verified numerically.

Keywords: Safety Certificates, Safe Control, Energy functions, Safety Index Synthesis, Constrained Reinforcement Learning, Multi-Objective Learning

1. Introduction

Safety is critical when applying state-of-the-art artificial intelligence studies to real-world applications, like autonomous driving (Sallab et al., 2017; Chen et al., 2021), robotic control (Richter et al., 2019; Ray et al., 2019). Safe control is one of the most common tasks among these real-world applications, requiring that the hard safety constraints must be obeyed persistently. However, learning

a safe control policy is hard for the naive trial-and-error mechanism of RL since it penalizes the dangerous actions *after* experiencing them.

Meanwhile, in the control theory, there exist studies about *energy-function-based* provable safety guarantee of dynamic systems called the *safety certificate*, or *safety index* (Wieland and Allgöwer, 2007; Ames et al., 2014; Liu and Tomizuka, 2014). These methods first synthesize energy functions such that the safe states have low energy, then design control laws satisfying the *safe action constraints* to make the systems dissipate energy (Wei and Liu, 2019). If there exists a feasible policy for *all states* in a safe set to satisfy the *safe action constraints* dissipating the energy, the system will never leave the safe set (i.e., forward invariance). Despite its soundness, the safety index synthesis (SIS) by hand is extremely hard for complex systems, which stimulates a rapidly growing interest in learning-based SIS (Chang et al., 2020; Saveriano and Lee, 2019; Srinivasan et al., 2020; Ma et al., 2021a; Qin et al., 2021). Nevertheless, These studies usually require known dynamical models (white-box, black-box or learning-calibrated) to design control laws. Furthermore, obtaining the policy satisfying safe action constraints is also challenging. Adding *safety shields or layers* to obtain supervised RL policies is a common approach (Wang et al., 2017; Agrawal and Sreenath, 2017; Cheng et al., 2019; Taylor et al., 2020), but these studies usually assume to know the valid safety certificates.

In general safe control tasks with unknown dynamics, one usually has access to *neither* the control laws *nor* perfect safety certificates, which makes the previous two kinds of studies fall into a paradox—they rely on each other as the prior knowledge. Therefore, this paper proposes a novel algorithm without prior knowledge about model-based control laws or valid safety certificates. We define a loss function for SIS by minimizing the occurrence of energy increases. Then we formulate a CRL problem (rather than the commonly used shield methods) to unify the loss functions of SIS and CRL. By adding SIS as an outer loop to the Lagrangian-based solution to CRL, we jointly update the policies and safety certificates, and prove that they will converge to their respective local optima, the optimal safe policies and the valid safety certificates.

Contributions. Our main contributions are: 1. We propose an algorithm of joint CRL and SIS that learns the safe policies and synthesizes the safety certificates simultaneously. This is the first algorithm requiring no prior knowledge of control laws or valid safety certificates. 2. We unify the loss function formulations of SIS and CRL. We therefore can form the multi-timescale adversarial RL training and prove its convergence. 3. We evaluate the proposed algorithm on multiple safety-critical benchmark environments Results demonstrate that we can simultaneously synthesize valid safety certificates and learn safe policies with zero constraint violation.

1.1. Related works

Representative energy-function-based safety certificates include barrier certificates (Prajna et al., 2007), control barrier functions (CBF) (Wieland and Allgöwer, 2007), safety set algorithm (SSA) (Liu and Tomizuka, 2014) and sliding mode methods (Gracia et al., 2013). Recent learning-based studies can be mainly divided into *learning-based SIS* and *learning safe control policies supervised by certificates*. Chang et al. (2020); Luo and Ma (2021) use explicit models to rollout or project actions to satisfy safe action constraints. Jin et al. (2020); Qin et al. (2021) guide certificate learning with LQR controllers, Zhao et al. (2021) requires a black-box model to query online, and Saveriano and Lee (2019); Srinivasan et al. (2020) use labeled data to fit certificates with supervised learning. The latter one, learning safe policy with supervisory usually assumed a valid safety certificate (Wang
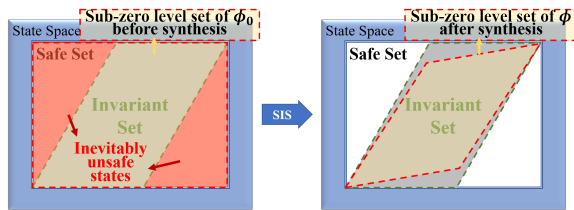
Figure 1: Safety index synthesis (SIS). Inevitably unsafe states should be excluded during the SIS.

et al., 2017; Cheng et al., 2019; Taylor et al., 2020). It's a natural thought that one could learn the dynamic models to handle these issues (like Cheng et al., 2019; Luo and Ma, 2021), but learning models is much more complex than only learning policies, especially in RL tasks.

2. Problem Formulations

We consider the safety specification that the system state s should be constrained in a connected and closed set $\mathcal{S}_s$ which is called the *safe set*. $\mathcal{S}_s$ should be a zero-sublevel set of a safety index function $\phi_0(\cdot)$ denoted by $\mathcal{S}_s = \{s|\phi_0(s) \leq 0\}$. We study the Markov decision process (MDP) with deterministic dynamics (a reasonable assumption when dealing with safe control problems), defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{F}, r, c, \gamma, \phi)$, where $\mathcal{S}, \mathcal{A}$ is the state and action space, $\mathcal{F} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the unknown system dynamics, $r, c : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward and cost function, γ is the discounted factor, and $\phi : \mathcal{S} \to \mathbb{R}$ is the energy-function-based safety certificate, or called the *safety index*. A safe control law with respect to safety index ϕ should keep the system energy low, ($\phi \leq 0$) and dissipate the energy when the system is at high energy ($\phi > 0$). We use s' to represent the next state for simplicity. Then we can get the *safe action constraint*:

Definition 1 (Safe action constraint) *For a given safety index ϕ, the safe action constraint is*

$$\phi(s') < \max\{\phi(s) - \eta_D, 0\} \tag{1}$$

where η_D is a slack variable controlling the descent rate of safety index.

Definition 2 (Valid safety certificate) *If there always exists an action $a \in \mathcal{A}$ satisfying (1) at s, or the safe action set $\mathcal{U}_s(s) = \{a|\phi(s') < \max\{\phi(s) - \eta_D, 0\}\}$ is always nonempty, we say the safety index ϕ is a **valid**, or **feasible** safety certificate.*

A straightforward approach is to use the ϕ_0 as the safety certificate. However, these safe action constraints are possibly not satisfied with all the states in $\mathcal{S}_s$ as shown in Figure 1. This problem is common in real-world tasks with actuator saturation and high relative-degree from safety specifications to control inputs (i.e., $\|\frac{\partial \phi_0}{\partial u}\| = 0$). For example, if ϕ_0 measures the distance between two autonomous vehicles, the collision may be inevitable because the relative speed is too high and brake force is limited. In this case, $\mathcal{S}_s$ includes inevitably unsafe states. We need to assign high energy values to these inevitably-unsafe states, for example, by linearly combining the ϕ_0 and its high-order derivatives (Liu and Tomizuka, 2014). The valid safety certificate will guarantee safety by ensuring the *forward invariance* of a subset of $\mathcal{S}_s$.

3

Lemma 3 (Forward invariance (Liu and Tomizuka, 2014)) *Define the zero-sublevel set of a valid safety index ϕ as $\mathcal{S}_s^\phi = \{s | \phi(s) \le 0\}$. If ϕ is a valid safety certificate, then there exist policies to guarantee the forward invariance of $\mathcal{S}_s^\phi \cap \mathcal{S}_s$.*

We therefore can formulate the CRL problem by adding the safe action constraints to RL optimization objective:

$$\max_\pi \ \mathbb{E}_{\tau \sim \pi} \Big\{ \sum\nolimits_{t=0}^\infty \gamma^t r_t \Big\} = \mathbb{E}_s \big\{ v^\pi(s) \big\} \quad \text{s.t. } \phi(s') - \max\{\phi(s) - \eta_D, 0\} < 0, \forall s \in \mathcal{D} \quad (2)$$

where $v^\pi(s)$ is the state-value function of s, $\mathcal{D} = \{s | f(s) > 0\}$ is the set of all possible states (f is the state distribution density).

Remark 4 (2) *has **state-dependent** constraints; it can not be solved by previous model-free CRL since their constraint objectives are defined on the **expectation** over $\mathcal{D}$ (Uchibe and Doya, 2007; Achiam et al., 2017; Chow et al., 2017; Tessler et al., 2018; Ray et al., 2019; Stooke et al., 2020).*

We solve (2) based on our previous framework to solve state-dependent safety constraints in (Ma et al., 2021b) using the Lagrange multiplier networks $\lambda(s)$ in a Lagrangian-based approach. The Lagrange function is (Ma et al., 2021b)

$$\mathcal{L}'(\pi, \lambda) = \mathbb{E}_s \big\{ -v^\pi(s) + \lambda(s) \big(\phi(s') - \max\{\phi(s) - \eta_D, 0\} \big) \big\} \quad (3)$$

We can solve (2) by locating the saddle point of $\mathcal{L}'(\pi, \lambda)$, $\max_\lambda \min_\pi \mathcal{L}'(\pi, \lambda)$.

3. Joint Synthesis of Safety Certificate and Safe Control Policy

The key idea of this section is to unify the loss functions of CRL and SIS; we provide theoretical analyses of their equivalence.

3.1. Loss Function for Safety Index Synthesis

We construct the loss for optimizing a parameterized safety index by a measurement of the *violation of constraint* (1)

$$J(\phi) = \mathbb{E}_s \Big\{ \big[\phi(s') - \max\{\phi(s) - \eta_D, 0\} \big]^+ \Big\} \Big|_{\pi=\pi^*} \quad (4)$$

where $[\cdot]^+$ means projecting the values to the positive half-space $[0, +\infty)$, π^* is the optimal safe policy (also a feasible policy when ϕ is a valid safety index) of (2), and $\cdot|_{\pi=\pi^*}$ represents the agent takes $\pi^*(s)$ to reach s'. Ideally, if ϕ is a valid safety index, there always exists control to satisfy (1), and $J(\phi) = 0$. For those imperfect ϕ, the inequality constraint in (2) may not hold for all states in $\mathcal{D}$, so we can optimize the loss to get better ϕ.

The joint synthesis algorithm is tricky since we need to handle *two different optimization problems*, (3) and (4). Recent similar studies integrate two optimizations by weighted sum (Qin et al., 2021) or alternative update (Luo and Ma, 2021), but their methods are more like intuitive approaches and lack a solid theoretical basis.

3.2. Unified Loss Function for Joint Synthesis

Lemma 5 (Statewise complementary slackness condition (Ma et al., 2021b)) *For the problem* (2), *if the safe action set is not empty at state* s, *the optimal multiplier and optimal policy* λ^*, π^* *satisfies*

$$\lambda^*(s)\{\phi(s') - \max\{\phi(s) - \eta_D, 0\}|_{\pi*}\} = 0 \tag{5}$$

If the safe action set is empty at state s, *then* $\lambda^*(s) \to \infty$.

The lemma comes from the Karush-Kuhn-Tucker (KKT) necessary conditions for the problem (2). Consider the Lagrange function (3) with the additional variable ϕ to optimize,

$$\mathcal{L}'(\pi, \lambda, \phi) = \mathbb{E}_s \left\{ -v^\pi(s) + \lambda(s)\big(\phi(s') - \max\{\phi(s) - \eta_D, 0\}\big) \right\} \tag{6}$$

we have the following lemma for the relationship between the loss function of policy and certificate synthesis

Lemma 6 *If* λ *is clipped into a compact set* $[0, \lambda_{\max}]$, *where* $\lambda_{\max} > \max_{s \in \{s | \mathcal{U}_s(s) \neq \emptyset\}} \lambda^*(s)$. *Then*

$$\mathcal{L}'(\pi^*, \lambda^*, \phi) = \lambda_{\max} J(\phi) + \Delta \tag{7}$$

where Δ *is a constant irrelevant with* ϕ.

Proof See Appendix A.1[1]. ∎

Theorem 7 (Unified loss for joint synthesis) *The optimal safety certificate parameters with optimal policy-multiplier tuple in* (6) *is also the optimal safety certificate parameters under loss in* (4)

$$\arg \min J(\phi) = \arg \min \mathcal{L}'(\pi^*, \lambda^*, \phi) \tag{8}$$

Proof See the *envelope theorem* on parameterized constraints in Milgrom and Segal (2002); Rockafellar (2015). ∎

Finally, we unify the loss function of updating three elements: policy π, multiplier λ, and safety index function ϕ. The optimization problem is formulated by a multi-timescale adversarial training:

$$\min_{\phi} \max_{\lambda} \min_{\pi} \mathcal{L}'(\pi, \lambda, \phi) \tag{9}$$

4. Practical Algorithm using Constrained Reinforcement Learning

In this section, we explain the practical algorithm and convergence analysis.

---

1. The full paper with Appendix can be found on https://arxiv.org/abs/2111.07695.

4.1. Algorithm Details

The Lagrangian-based solution to CRL with statewise safety constraint is compatible with any existing unconstrained RL baselines, and we select the off-policy maximum entropy RL framework like soft actor-critic (SAC, Haarnoja et al., 2018). According to (3), we need to add two neural networks to learn a state-action value function for safety index model $Q_\phi(s, a)$ (to approximate $\phi(s') - \max\{\phi(s) - \eta_D, 0\}$) and the multipliers $\lambda(s)$. Because the soft policy evaluation, or the soft Q-learning part, is the same as Haarnoja et al. (2018), we only demonstrate the soft policy iteration part of the algorithm in Algorithm 1. We name the proposed algorithm as FAC-SIS[2]. We denote the parameters of the policy network, multiplier network, safety index as θ, ξ, ζ, respectively. We use $G_\#$, $\# \in \{\theta, \xi, \zeta\}$ to denote the gradients to update policy, multiplier and certificate. Detailed computations of the gradients can be found in Appendix B.1. In addition, we assign multiple delayed updates (similar to Fujimoto et al., 2018), $m_\pi < m_\lambda < m_\phi$, to stabilize the adversarial optimizations.

Remark 8 *A general parameterization rule of safety index is to linearly combine ϕ_0 and its high-order derivatives (Liu and Tomizuka, 2014), $\phi = \phi_0 + k_1\dot{\phi}_0 + \cdots + k_n\phi_0^{(n)}$; the parameters ξ is $[k_1, k_2, \ldots, k_n]$. How many high-order derivatives are needed depends on the system relative-degree. For example, the relative-degree of position constraints with force inputs is 2. This information should be included in observations of MDP. Otherwise, the observation can not fully describe how dangerous the agent is with respect to the safety constraint.*

Algorithm 1 Soft Policy Improvement in FAC-SIS

Input: Buffer $\mathcal{D}$ with sampled data, policy parameters θ, multiplier parameters ξ, safety index parameters ζ.
  1: **if** gradient steps mod $m_\pi = 0$ **then** $\theta \leftarrow \theta - \overline{\beta_\pi}G_\theta$
  2: **if** gradient steps mod $m_\lambda = 0$ **then** $\xi \leftarrow \xi + \overline{\beta_\lambda}G_\xi$
  3: **if** gradient steps mod $m_\phi = 0$ **then** $\zeta \leftarrow \zeta - \overline{\beta_\zeta}G_\zeta$
Output: w_1, w_2, θ, ξ.

4.2. Convergence Analysis

The convergence proof of a three timescale adversarial training of (9) mainly follows the multi-timescale convergence according to Theorem 2 in Chapter 6 in Borkar (2009) about multiple timescale convergence of multi-variable optimization. Some studies also adopted this procedure to explain the convergence of RL algorithms from the perspective of stochastic optimization (Bhatnagar et al., 2009; Bhatnagar and Lakshmanan, 2012), especially those with Lagrangian-based methods (Chow et al., 2017). We incorporate the recent study on *clipped stochastic gradient descent* to further improve the generalization of this convergence proof (Zhang et al., 2019). We first give some assumptions:

Assumption 1 (learning rates) *The learning rate schedules, $\{\beta_\theta(k), \beta_\xi(k), \beta_\zeta(k)\}$, satisfy*

$$\sum_k \beta_\theta(k) = \sum_k \beta_\xi(k) = \sum_k \beta_\zeta(k) = \infty.$$

$$\sum_k \beta_\theta(k)^2, \quad \sum_k \beta_\xi(k)^2, \quad \sum_k \beta_\zeta(k)^2 < \infty, \quad \beta_\xi(k) = o\left(\beta_\theta(k)\right), \, \beta_\zeta(k) = o\left(\beta_\xi(k)\right).$$

2. FAC refers to the FAC algorithm in our prior work(Ma et al., 2021b).

This assumption also implies that the policy converges in the fastest timescale, then the multipliers, and finally the safety index parameters.

Proposition 9 (Clipped gradient descent) *The actual learning rate used in Algorithm 1 is*

$$\overline{\beta_\#} := \min\{\beta_\#, \beta_\#/\|G_\#\|\} \tag{10}$$

where $\# \in \{w_1, w_2, \theta, \xi, \zeta\}$ is the parameters, and $G_\#$ is the corresponding gradients.

Assumption 2 *The state and action are sampled from compact sets, and all neural networks are $L_0 - L_1$ smooth.*

As we are to finish a safe control problem that the agent should be confined in safe sets, and the actuator has physical limits, the bounded assumption is reasonable. We use multi-layer perceptron with continuous differentiable activation functions in practical implementations (details can be found in Appendix D).

Theorem 10 *Under all the aforementioned assumptions, the sequence of policy, multiplier, and safety index parameters tuple $(\theta_k, \xi_k, \zeta_k)$ converge almost surely to a locally optimal safety index parameters and its corresponding locally optimal policy and multiplier $(\theta^*, \xi^*, \zeta^*)$ as k goes to infinity.*

Proof See Appendix B.2. ∎

5. Experiments

In our experiments, we focus on the following questions:

1. How does the proposed algorithm compare with other constraint RL algorithms? Can it achieve a safe policy with zero constraint violation?

2. How does the learning-based safety index synthesis outperform the handcrafted safety index or the original safety index in the safety performance?

3. Does the synthesized safety index allow safe control in all states the agent experienced?

To demonstrate the effectiveness of the proposed online synthesis rules, we select the safe reinforcement learning benchmark environments Safety Gym (Ray et al., 2019) with different tasks and obstacles. We name a specific environment by {Obstacle type}–{Obstacle size}–{Task}. We select six environments with different tasks and constraint objectives, where four of them are demonstrated in Figure 2, and others are provided in Appendix E.1.

Remark 11 *Some previous studies use controllers specially designed for goal-reaching tasks (Jin et al., 2020; Qin et al., 2021), while our algorithm can handle arbitrary complex tasks like the* Push *tasks.*

In this section, we use a fine-tuned form of safety index in Zhao et al. (2021)

$$\phi(s) = \sigma + d_{\min}^n - d^n - k\dot{d} \tag{11}$$

where d is the distance between the agent and obstacle, $d_{\min}$ is the minimum safe distance, and $\dot{d}$ is the derivative of distance with respect to time, $\xi = [\sigma, n, k]$ are the tunable parameters we desire to optimize in the online synthesis algorithm. The observations in Safety Gym include Lidar, speedometer, and magnetometer, which can be used to compute d and $\dot{d}$ from observations. We compare the proposed algorithm against two types of baseline algorithms:

- CRL baselines, including TRPO-Lagrangian, PPO-Lagrangian and CPO Achiam et al. (2017); Ray et al. (2019). The cost threshold is set at zero to learn solid safe policies.

- FAC with original safety index ϕ_0 and handcrafted safety index ϕ_h, where $\phi_0 = d_{min} - d$ and $\phi_h = 0.3 + d_{min}^2 - d^2 - k\dot{d}$, named as *FAC with ϕ_0* and *FAC with ϕ_h*. The choice of ϕ_h is based on empirical knowledge. Details about baseline algorithms can be found in Appendix C.1.

5.1. Evaluating FAC-SIS and Comparison Analysis

Results of the performance comparison are shown in Figure 3. The results suggest that FAC with ϕ_0 performs poorest in the safety performance, which indicates that there are indeed many inevitably unsafe states, and SIS is necessary for these tasks. Only FAC-SIS learns the policy with *zero violation* and takes the lowest cost rate in all environments, answering the first question of zero constraint violation. FAC with ϕ_h fails to learn a solid safe policy in those environments with `0.30` constraint size (See the zoomed window in the second row), indicating that the handcrafted safety index can not cover all the environments. As for the baseline CRL algorithms, they can not learn a zero-violation policy in any environment because of the posterior penalty in the trial-and-error mechanism stated above. As for the reward performance, FAC-SIS has comparable reward performance in the `Push` task. For the `Goal` task, FAC with ϕ_h and FAC-SIS sacrifice the reward performance to guarantee safety, explained in Ray et al. (2019).

5.2. Validity Verification of Synthesized Safety Index

We conduct new metrics and experiments to show the validity of our synthesized safety index. Recall that the validity means that there always exists a feasible control policy to satisfy the constraint

(a) Hazards: non-physical dangerous areas. (b) Pillars: Static physical obstacles (c) Goal task: navigating the robot to reach the green cylinder. (d) Push task: pushing the yellow box inside the green cylinder.
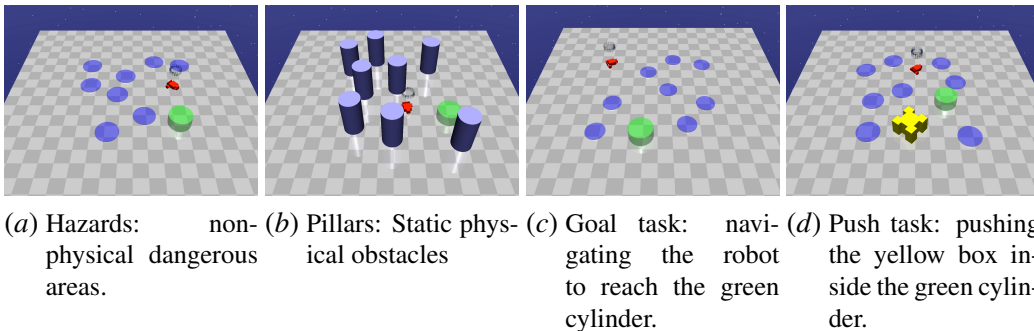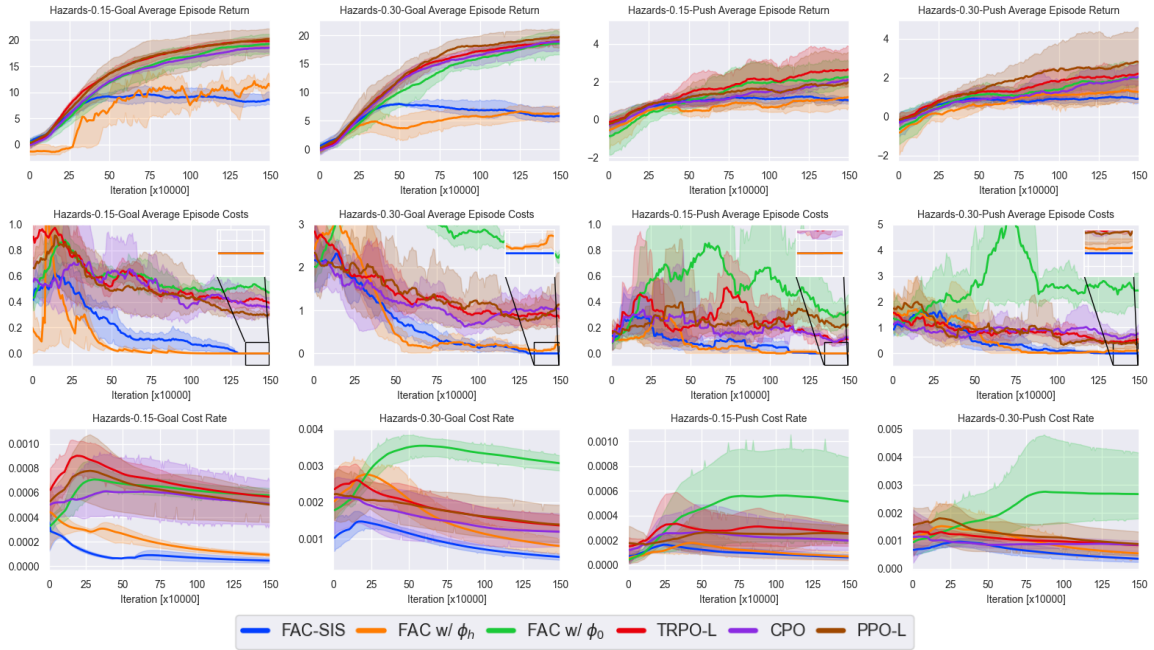
Figure 2: Obstacles and tasks in Safety Gym.

Figure 3: Average performance of FAC-SIS and baseline methods on 4 different Safety Gym environments over five seeds.
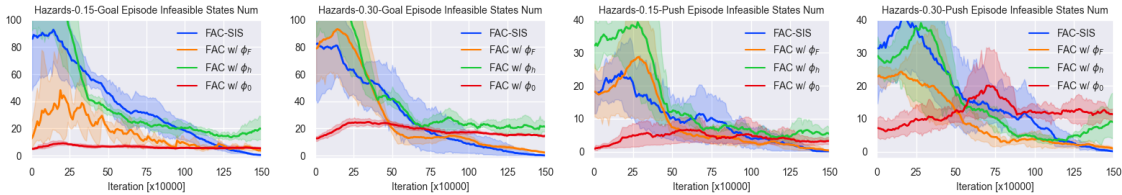

Figure 4: Average episodic number of violations of safe action constraint (1). A valid safety index and its corresponding safe control policy should have zero violation performance.

in (1). To effectively demonstrate the feasibility of SIS, we add a new baseline, *FAC with* ϕ_F, where ϕ_F is a valid safety index verified by Zhao et al. (2021). Figure 4 demonstrates the episodic number of constraint violations of (1) in the Safety Gym environments. The results show that FAC-SIS and *FAC with* ϕ_F can reach a nearly stable zero violation, which means that FAC can satisfy safe action constraint for a given valid safety index, and FAC-SIS also synthesizes a valid safety index. However, with ϕ_0 and ϕ_h there are consistent violations even with the converged policy, caused by different reasons. For ϕ_h, the reason is simply the inability to make the energy dissipate. For ϕ_0, no high-order derivative in the safety index, so ϕ_0 cannot handle the high relative-degree between the constraint function and control input.

Furthermore, we want to give some scalable analyses about SIS. Firstly, we want to visualize that we can always find actions to dissipate the energy in the sampled distributions after SIS. We
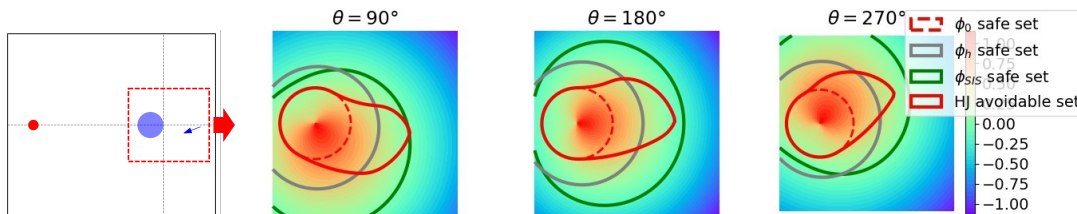
Figure 5: The color indicates the value of the synthesized safety index. The zero-sublevel set of learned safety index is a subset of the numerically max forward invariant set, the HJ avoidable set. We exclude all inevitably unsafe states through SIS like Figure 1. HJ-based solution is only capable for goal-reaching tasks with the reach-avoid problem formulation, while proposed algorithm can handle arbitrary task.

conduct experiments on a simple collision-avoidance environment shown in Figure **??**[3]. We select three different initialization rules of the agent and hazard. We use a sample-based exhaustive method to identify infeasible states and see if they overlap the sampled state distributions. We project the state distribution to the 2D space of d and $\dot{d}$, and the results are listed in Figure **??**. As we expect all the sampled states to be feasible, these two distributions *should not overlap*. The results show that the overlap of these two distributions is very small, which indicates that nearly all the sampling states are feasible.

Secondly, we want to visualize the shape of the safe set concerning the learned safety index. We slice the state space into three 2D planes with different agent heading angles, as shown in Figure 5. We use Hamilton-Jacobi reachability analysis to compute the avoidable sets numerically. The avoidable set considers the safe set under the *most conservative* control inputs, which is the maximum controlled invariant set (Mitchell, 2007; Choi et al., 2021). We use ϕ_h as the initial safety index of SIS. Figure 5 demonstrates that inevitable unsafe states exist in the zero-sublevel set of the empirical safety index. It also shows that we successfully exclude the inevitably unsafe sets through SIS. Notably, the zero-sublevel sets of the synthesized safety index are the subsets of the HJ-avoidable sets. The reasons why SIS can not learn the perfect shapes include limits of the representation capabilities of the safety index parameterization. Additionally, we still consider the *optimal criteria*, resulting in less conservative policies and possibly smaller safe sets.

6. Conclusion

This paper focuses on the joint synthesis of the safety certificate and the safe control policy for unknown dynamical systems and general tasks using a CRL approach. We are the first study to start with unknown dynamics and imperfect safety certificates, which significantly improves the applicability of the energy-function-based safe control. We add the optimization of safety index parameters as an outer loop of Lagrangian-based CRL with a unified loss. The convergence is analyzed theoretically. Experimental results demonstrate that the proposed FAC-SIS synthesizes a valid safe index while learning a safe control policy. In future work, we will consider more complex safety index parameterization rules, for example, neural networks. Meanwhile, we will consider other factors in SIS, such as the reward performance of the safe control policies.

---

3. See Appendix E.2 for more details.

Acknowledgments

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31, Sydney, Australia, 2017. PMLR.

Ayush Agrawal and Koushil Sreenath. Discrete control barrier functions for safety-critical control of discrete systems with application to bipedal robot navigation. In *Robotics: Science and Systems*, 2017.

Aaron D Ames, Jessy W Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs with application to adaptive cruise control. In *53rd IEEE Conference on Decision and Control*, pages 6271–6278. IEEE, 2014.

Shalabh Bhatnagar and K Lakshmanan. An online actor–critic algorithm with function approximation for constrained markov decision processes. *Journal of Optimization Theory and Applications*, 153(3):688–708, 2012.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Ya-Chien Chang, Nima Roohi, and Sicun Gao. Neural lyapunov control. *arXiv preprint arXiv:2005.00611*, 2020.

Jianyu Chen, Shengbo Eben Li, and Masayoshi Tomizuka. Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Richard Cheng, Gábor Orosz, Richard M Murray, and Joel W Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3387–3395, 2019.

Jason J Choi, Donggun Lee, Koushil Sreenath, Claire J Tomlin, and Sylvia L Herbert. Robust control barrier-value functions for safety-critical control. *arXiv preprint arXiv:2104.02808*, 2021.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pages 1587–1596, Stockholm, Sweden, 2018. PMLR.

Luis Gracia, Fabricio Garelli, and Antonio Sala. Reactive sliding-mode algorithm for collision avoidance in robotic systems. *IEEE Transactions on Control Systems Technology*, 21(6):2391–2399, 2013.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870, Stockholm, Sweden, 2018. PMLR.

Wanxin Jin, Zhaoran Wang, Zhuoran Yang, and Shaoshuai Mou. Neural certificates for safe control policies. *arXiv preprint arXiv:2006.08465*, 2020.

Changliu Liu and Masayoshi Tomizuka. Control in a safe set: Addressing safety in human-robot interactions. In *Dynamic Systems and Control Conference*, volume 46209, page V003T42A003. American Society of Mechanical Engineers, 2014.

Yuping Luo and Tengyu Ma. Learning barrier certificates: Towards safe reinforcement learning with zero training-time violations. *arXiv preprint arXiv:2108.01846*, 2021.

Haitong Ma, Jianyu Chen, Shengbo Eben Li, Ziyu Lin, and Sifa Zheng. Model-based constrained reinforcement learning using generalized control barrier function. *arXiv preprint arXiv:2103.01556*, 2021a.

Haitong Ma, Yang Guan, Shegnbo Eben Li, Xiangteng Zhang, Sifa Zheng, and Jianyu Chen. Feasible actor-critic: Constrained reinforcement learning for ensuring statewise safety. *arXiv preprint arXiv:2105.10682*, 2021b.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

Ian M Mitchell. A toolbox of level set methods. *UBC Department of Computer Science Technical Report TR-2007-11*, 2007.

Stephen Prajna, Ali Jadbabaie, and George J Pappas. A framework for worst-case and stochastic safety verification using barrier certificates. *IEEE Transactions on Automatic Control*, 52(8):1415–1428, 2007.

Zengyi Qin, Kaiqing Zhang, Yuxiao Chen, Jingkai Chen, and Chuchu Fan. Learning safe multi-agent control with decentralized neural barrier certificates. *arXiv preprint arXiv:2101.05436*, 2021.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 2019.

Florian Richter, Ryan K Orosco, and Michael C Yip. Open-sourced reinforcement learning environments for surgical robotics. *arXiv preprint arXiv:1903.02090*, 2019.

Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017.

Matteo Saveriano and Dongheui Lee. Learning barrier functions for constrained motion planning with dynamical systems. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 112–119. IEEE, 2019.

Mohit Srinivasan, Amogh Dabholkar, Samuel Coogan, and Patricio A Vela. Synthesis of control barrier functions using a supervised machine learning approach. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7139–7145. IEEE, 2020.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pages 9133–9143, Online, 2020. PMLR.

Andrew Taylor, Andrew Singletary, Yisong Yue, and Aaron Ames. Learning for safety-critical control with control barrier functions. In *Learning for Dynamics and Control*, pages 708–717. PMLR, 2020.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.

Eiji Uchibe and Kenji Doya. Constrained reinforcement learning from intrinsic and extrinsic rewards. In *2007 IEEE 6th International Conference on Development and Learning*, pages 163–168, Lugano, Switzerland, 2007. IEEE.

Li Wang, Aaron D Ames, and Magnus Egerstedt. Safety barrier certificates for collisions-free multirobot systems. *IEEE Transactions on Robotics*, 33(3):661–674, 2017.

Tianhao Wei and Changliu Liu. Safe control algorithms using energy functions: A uni ed framework, benchmark, and new directions. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 238–243. IEEE, 2019.

Peter Wieland and Frank Allgöwer. Constructive safety using control barrier functions. *IFAC Proceedings Volumes*, 40(12):462–467, 2007.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

Weiye Zhao, Tairan He, and Changliu Liu. Model-free safe control for zero-violation reinforcement learning. In *5th Annual Conference on Robot Learning*, 2021.