Input-to-State Safety for Reinforcement Learning

Mayank Shekhar Jha[1], Satya Marthi[1], Kyriakos G. Vamvoudakis[2], Soha Kanso[1], Didier Theilliol[1]

***Abstract*—In this paper, we present a novel off-policy, safe reinforcement learning approach for nonlinear dynamical systems under input saturation, that guarantees safe initialization, safe exploration as well as safe learning of optimal control laws. First, to encourage preferable exploration near safety boundaries, important for integrating system behavior near the safety limits; we formulate a safe exploration approach as a robust control problem by considering an enlarged safe set based on Input-to-State Safe Control Barrier Functions (ISSf-CBF). These constraints are then incorporated into a quadratic programming (QP) optimization. We propose a novel ϵ-tuning law that adaptively enforces stricter safety constraints near the boundaries of the safe set and relaxes constraints deeper within the safe set, encouraging safety boundary-proximal exploration while maintaining forward invariance of the safe set. The proposed ϵ-tuning law safely accommodates aggressive, high-magnitude exploration noise, enabling efficient state-space exploration without compromising safety. Next, safe learning under saturation limits is guaranteed through safety aware cost function. We establish safety, optimality and stability properties (novel) in a mathematically rigorous manner. Further, the safe RL problem is solved in an off-policy manner, and neural networks are used to approximate the value function and the control policy. To that end, we establish novel off-policy equation under input saturation. Finally, simulations demonstrate the efficacy of the proposed framework.**

Keywords: Safe reinforcement learning, neural networks, input to state safety, safe control learning.

I. Introduction

Research efforts have focused on designing and learning control laws for dynamical systems that ensure both system safety and stability while also meeting performance requirements [5, 35]. Safety can be formally characterized through the forward invariance of prescribed sets within the state space, which can be achieved using barrier certificates, barrier functions [36] and control barrier functions (CBF) [4]. Two types of CBFs have been introduced in the literature: Zeroing CBF (ZCBF) and Reciprocal CBF (RCBF) [4]. Although both serve the same objective, RCBFs remain unbounded on the boundary, and ZCBFs typically vanish on the boundary [6, 20].

However, unmodeled dynamics, uncertainties, and disturbances require the development of robust approaches to ensure safety in practical applications. To address these issues, the concept of input-to-state-safety (ISSf) was introduced in [28] and later extended to address bounded disturbances in [15], presenting a comprehensive framework where safety is guaranteed by ensuring forward invariance of a larger set, leading to the formalization of input-to-state safe control barrier functions (ISSf-CBF). However, ISSf-CBF based approaches limit design flexibility, rendering control laws conservative. This was recently addressed in [2, 3] with tunable ISSf CBF (TISSf-CBF). Specifically, a generalized version of ISSf-CBF was proposed that provided the means of tuning the size of the larger invariant set so that it approximates the safe set of the undisturbed system without having a significant impact on performance, leading to a reduction in conservatism. On the other hand, Control Lyapunov Functions (CLF) are typically used to characterize the family of controllers that stabilize the system [30]. CLF and CBF have been unified within a Quadratic Programming (QP)-based control optimization framework, allowing for the satisfaction of stability and safety for optimal regulation and optimal tracking cases [4, 9].

Reinforcement learning (RL) based approaches learn optimal control policies/laws while guaranteeing system stability and optimality [16]. Systems under input constraints have also been considered in previous works: [1, 17], to solve Hamilton-Jacobi-Bellman (HJB) equation with non-quadratic functional under input constraints. In [23] it was extended to completely unknown systems and [14] addressed optimal tracking problem. However, it remains unclear how to consider such a non-quadratic performance functional under safety considerations and solve the corresponding HJB for systems under input constraints.

RL operates in two phases: exploration and exploitation [11]. Although the exploration phase primarily involves data collection under noisy inputs such as random noise or exponentially decreasing probing noise [33], the exploitation phase involves learning the optimal policy using collected data. Exploration noise can lead to instability or safety violations as the system visits unsafe regions [12].

Safe RL has recently emerged as a promising research domain that aims to guarantee the safety of the system along with stability and optimal performance [5]. Although a variety of approaches have been developed, including modification of the worst-case and risk-sensitive criterion [31], this work emphasizes safe RL based on CBF in the sense of [12, 19] to build control-theoretic centric approaches for non-linear systems. Safe RL has witnessed a surge in research activities recently [10, 12, 19]. In CBF-based context, safety is typically achieved by assuring invariance of a prescribed safe set as system-states evolve over time [10, 12, 19]. In particular, [19] developed the Safe RL approach by augmenting the traditional cost function with RCBF, leading to safety guarantees during the learning phase and [10] extended the approach for nonlinear discrete-time systems. From a methodological

[1]M. S. Jha, S. Marthi, and D. Theilliol are with CRAN, CNRS, University of Lorraine, Nancy, France 54000. Email: mayank-shekhar.jha@univ-lorraine.fr

[2]K. G. Vamvoudakis is with the Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Email: kyriakos@gatech.edu.

This material was based upon work supported in part by French National Research Project ANR SOS (ANR-23-IAS2-0003), by NSF under grant No. SLES-2415479, by NASA ULI under grant No. 80NSSC25M7104, and by ARO under grant No. W911NF-24 − 1 − 0174.

perspective, safe RL has progressed from encoding safety as constraints and/or risk-sensitive robust criteria within the learning objective (e.g., worst-case formulations) [31], toward control-theoretic mechanisms that enforce forward invariance through online safety filters based on Lyapunov and barrier certificates (e.g., CLF/CBF-QP shielding) [4, 8, 12, 19]. More recently, robust barrier constructions based on input-to-state safety guarantee invariance under bounded disturbances by enlarging the admissible set [15], while tunable ISSf-CBFs reduce conservatism by explicitly parameterizing the enlargement [2, 3]. This paper builds on these developments to enable boundary-aware safe exploration under persistent excitation and actuator saturation within an off-policy learning architecture.

However, it becomes imperative to ensure that system states remain within the safe set during the exploration phase (data collection) as well as the exploitation phase in a holistic sense. In these aforementioned works, safety during the exploration phase has been completely neglected. Moreover, the knowledge of an initial safe and admissible control policy, essential for initialization of iterative control learning, is assumed. This assumption was removed in [12], wherein CLF and RCBF conditions were unified as linear constraints within a QP-based optimization problem, enforcing stabilization and safety on the controller performance, respectively, leading to safe exploration guarantees and the so-called end-to-end safety, i.e., safe initialization, exploration (data collection), and learning (exploitation). However, the latter did not consider constraints on control inputs (actuator saturation).

Finally, it is important to note that generally, policy iteration (PI) algorithms can be implemented in two distinct manners: on-policy and off-policy [11, 19]. Off-policy-based approaches are preferred over on-policy ones for safe RL as they involve the usage of the so-called behavioral policy for data collection (exploration phase) and a separate policy called the target policy for improvement towards an optimal one during the learning phase [10, 12, 19]. However, most of the existing aforementioned works have not developed off-policy expression for Safe RL under actuator saturation.

Motivations: During exploration, informative data are typically generated when trajectories evolve close to safety boundaries, since the closed-loop response in such regimes determines whether the learned policy can safely handle near-critical operating conditions. However, achieving persistent excitation in continuous-time systems often requires high-magnitude probing signals, which can make standard CBF-based exploration overly conservative (or even infeasible) near boundary, thereby degrading data quality exactly where it is most needed. None of the existing works focus on this problem. Moreover, actuator saturation is a fundamental physical limitation that couples safety and learning: the applied behavior input is clipped, which alters the generated trajectories and can invalidate off-policy learning updates if saturation is ignored. None of the existing works consider input saturation, as well as the latter under the off-policy scheme; motivating the present work.

Contributions: The contributions of the present work are fourfold.

First, we formulate safe exploration as a robust safety problem by introducing an ISSf-type enlarged set and enforcing the resulting ISSf-Exp-CBF constraint through a CLF/CBF quadratic program with actuator saturation bounds.

Second, we propose an exponential $\epsilon(h)$ tuning law that tightens the exploration constraint near the boundary but relaxes it within the interior of safe set, reducing conservatism while enabling informative, boundary-aware exploration under high-magnitude persistent excitation.

Third, we incorporate input saturation consistently in both the online safety filter and the safety-aware learning objective, and we rigorously establish end-to-end safety, stability, and optimality guarantees.

Finally, for implementation we use an off-policy approach and to that end, we establish rigorously a novel closed-form analytical expression to solve the proposed approach.

Structure: The remainder of the paper is structured as follows. Section II provides the background and presents the problem under consideration. Section III introduces a novel safe exploration approach. Section IV proposes a novel safe learning approach under input saturation, while Section V presents a novel off-policy RL approach. Section VI presents simulation results. Finally, Section VII concludes and suggests future directions.

Notation: $\mathcal{C}^1$ represents the set of continuously differentiable functions. The interior and boundary of $\mathcal{C}$, respectively, are denoted by $\mathrm{Int}(\mathcal{C})$ and $\partial\mathcal{C}$. For any essentially bounded function $d : \mathbb{R} \to \mathbb{R}^n$, its infinity norm is denoted by $\|d\|_\infty = \operatorname{ess\,sup}_{t\in\mathbb{R}} \|d(t)\|$. A continuous function $\alpha_1 : [0, a) \to [0, \infty)$ for some $a > 0$ is said to belong to class $\mathcal{K}$ if it is strictly increasing and $\alpha_1(0) = 0$. Similarly, a continuous function $\alpha : [0, \infty) \to [0, \infty)$ is said to be class $\mathcal{K}_\infty$ $(\alpha \in \mathcal{K}_\infty)$ if it is strictly monotonically increasing with $\alpha(0) = 0$ and $\lim_{r\to\infty} \alpha(r) = \infty$. A continuous function $\alpha : \mathbb{R} \to \mathbb{R}$ belongs to extended class $\mathcal{K}^e_\infty$ $(\alpha \in \mathcal{K}^e_\infty)$ if $\alpha(0) = 0, \alpha$ is strictly increasing, and $\lim_{r\to\infty} \alpha(r) = \infty$ and $\lim_{r\to-\infty} \alpha(r) = -\infty$.

II. Recent related works

[41] learns a low-dimensional safe-region representation to improve scalability, but does not address robustness of persistently exciting exploration or actuator saturation. [22] develops a model-free safety certifier, yet without boundary-aware ISSf-type set enlargement or saturation-aware policy-iteration updates.[40] integrates Barrier Lyapunov Functions with RL via optimized back-stepping, emphasizing constructive safe stabilization rather than tunable disturbance-to-safe-set robustness for aggressive exploration. Event-triggered safe RL has also been studied, focusing on triggering/resource mechanisms rather than adaptive tightening/relaxation of exploration constraints near the boundary [39]. [34] propose shielded planning guided policy optimization, different from the proposed CBF/QP-based robust invariance analysis under bounded exploration. In robotics, model-free neural barrier certificates have also been used as safety constraints during RL training, though without the ISSf-style tunable enlargement used to safely accommodate high-magnitude excitation

[37]. [38] develops safety-optimal fault-tolerant control (FTC) via adaptive critic design with asymmetric input constraints, focusing on FTC objectives rather than boundary-aware exploration shaping. [25] studies RL-based safe tracking with event-triggered updates, emphasizing security and resource-aware execution rather than exploration filtering near safety boundaries. [27] proposes event-triggered observer-based FTC for saturated nonlinear systems with state constraints, but does not integrate a CLF/CBF–QP exploration filter nor an off-policy least-squares policy-iteration update under saturation. Finally, [26] introduces a barrier-critic robust framework for constrained differential games, which targets game-theoretic robustness rather than end-to-end safe RL. In contrast, we develop an end-to-end safe RL framework with tunable input-to-state safe exploration, a CLF/CBF–QP safety filter enforcing saturation during exploration, and an off-policy least-squares policy-iteration update under saturation.

III. Background and Problem Formulation

Consider an affine in the control nonlinear dynamical system under actuator saturation, $\forall t \geq 0$, as given by the equation:

$$\dot{x} = f(x) + g(x)u(t) \tag{1}$$

where, $x \in \Omega \subseteq \mathbb{R}^n$ represents the state of the system over a compact set Ω. The control input satisfies actuator saturation constraint $u(t) \in \mathcal{U} \triangleq \left\{ u \in \mathbb{R}^m \;\middle|\; |u_i| \leq \rho,\ i = 1, \ldots, m \right\}$ or $\|u\|_\infty \leq \rho$ where $\rho > 0$ is a known saturation bound, $f(x) \in \mathbb{R}^n$ and $g(x) \in \mathbb{R}^{n \times m}$ are the drift and the input dynamics of the system, respectively. Without loss of generality, $x = 0$ is an equilibrium state such that $f(0) = 0$ and $g(0)$ is well defined. It is also assumed that system (1) is stabilizable on $\Omega \subseteq \mathbb{R}^n$.

A. Optimal Control under Saturation

Define a generalized non-quadratic functional performance index as:

$$V(x, u) = \int_t^\infty (Q(x(\tau)) + U(u(\tau)))\mathrm{d}\tau \tag{2}$$

where $Q(x)$ is a positive-definite monotonically increasing function, and $U(u)$ is a positive definite integrand function.

Assumption 1 (Zero state observability [23]). The performance function (2) satisfies the zero-state observability property.

Definition 1 (Admissibility [1])**.** *A sequence of control inputs/policies, denoted by $u(x) \in \pi(\Omega)$, is considered admissible with respect to (2) in Ω if $u(x)$ is continuous on Ω, $u(0) = 0$, and stabilizes the system (1) on Ω. Additionally, $V(x)$ must be finite for all $x \in \Omega$.*

To find an admissible control policy $u(x) : \mathbb{R}^n \to \mathbb{R}^m$ that minimizes the generalized performance index (2), we must ensure that the control sequence inputs are saturated, i.e., $u(x) \in \mathcal{U}$. To accomplish this, we employ a generalized non-quadratic function as specified in [23],

$$U(u) = 2 \int_0^u \rho \left(\Gamma^{-1}(\frac{v}{\rho}) \right)^\top R\mathrm{d}v \tag{3}$$

where $\Gamma(\cdot) = \tanh(\cdot)$ and $v \in \mathbb{R}^m$ and $R = \mathrm{diag}\,(r_1, \ldots, r_m) > 0$. The control input bounds can be scaled by ρ since $|\tanh(\cdot)| \leq 1$ and $\tanh(0) = 0$. Furthermore, $\tanh(\cdot)$ is a monotonic odd function, and its first derivative is bounded by a constant. It is noted that in (3), the integral is understood component-wise (separable), i.e., $U(u) = 2\sum_{j=1}^m \int_0^{u_j} \rho\, r_j\, \Gamma^{-1}\left(\frac{v}{\rho}\right) dv$, with Γ^{-1} applied element-wise. Taking the partial derivative of the value function $V(\cdot)$ along the system (1) trajectories, the Generalized Hamilton-Jacobi-Bellman equation (G-HJB) [1, 23] is obtained as

$$\begin{aligned} &\mathrm{GHJB}(V, u) = \nabla V^\top(x)(f(x) + g(x)u(x)) \\ &+ Q(x) + 2\int_0^u \left(\rho \tanh^{-1}(\frac{v}{\rho}) \right)^\top R\mathrm{d}v = 0 \\ &V(0) = 0 \end{aligned} \tag{4}$$

where $\nabla V(x) = \partial V(x)/\partial x$. Denote $V^\star(x)$ as the optimal cost function given $\forall x$ by

$$V^\star(x(t)) = \min_{\substack{u(\tau) \in \pi(\Omega) \\ t \leq \tau < \infty}} \int_t^\infty (Q(x(\tau)) + U(u(\tau)))\mathrm{d}\tau. \tag{5}$$

Then, by substituting $V^\star(x)$ in (4), the G-HJB $(V, u) = 0$ becomes

$$\begin{aligned} &H\left(x, u, \nabla V^\star\right): \\ &\min_{u \in \pi(\Omega)} \Bigg[Q(x) + 2\int_0^u \left(\rho \tanh^{-1}(\frac{v}{\rho}) \right)^\top R\mathrm{d}v \\ &+ \nabla V^{\star\top}(x)(f(x) + g(x)u(x)) \Big] = 0, \\ &V^\star(0) = 0. \end{aligned} \tag{6}$$

with an optimal control given $\forall x$ by,

$$u^\star(x) = -\rho \tanh\left(\frac{1}{2\rho} R^{-1} g^\top(x) \nabla V^\star(x) \right). \tag{7}$$

Substituting (7) into (4) leads to the HJB [1] that does not have a closed-form solution. It should be noted that asymmetric saturation can be handled by replacing the symmetric bounds in U with component-wise limits $u_{\min,i} \leq u_i \leq u_{\max,i}$ and enforcing these as constraints in the QP-based safety filter.

B. Safety

The concept of safety is now established using a safe set $\mathcal{C} \subseteq \Omega \in \mathbb{R}^n$ that must be forward invariant. This implies that as the system states evolve according to (1) [4, 9, 19], the safe set remains unchanged. Consider a user-defined safety function h that is selected to encode the user-defined safety specification (e.g., box/polytope/ellipsoid or distance-to-obstacle constraints admit analytic forms), that maps $\mathbb{R}^n$ to $\mathbb{R}$ and is smooth, which means it belongs to the class C^1. The safe set is then defined as follows:

$$\mathcal{C} = \{x \in \mathbb{R}^n : h(x) \geq 0\} \tag{8}$$

$$\partial\mathcal{C} = \{x \in \mathbb{R}^n : h(x) = 0\} \tag{9}$$

$$\mathrm{Int}(\mathcal{C}) = \{x \in \mathbb{R}^n : h(x) > 0\} \tag{10}$$

where, $\partial\mathcal{C}$ and $\mathrm{Int}(\mathcal{C})$ represent the boundary and the interior of the set $\mathcal{C}$, respectively. The following definitions are needed before we proceed.

Definition 2 (Forward Invariance & Safety)**.** *A set $\mathcal{C} \subset \mathbb{R}^n$ is forward invariant if, for every $x_0 \in \mathcal{C}$, the solution $x(t)$ to the system (1) satisfies $x(t) \in \mathcal{C}$ for all $t \geq 0$. In other words, the system (1) is safe with respect to the set $\mathcal{C}$ if and only if $\mathcal{C}$ is forward invariant.*

Two types of Barrier Functions (BF) are developed in the literature: Reciprocal Control Barrier Functions (RCBFs) that become unbounded at the safe set boundary and Zeroing CBFs (ZCBFs), that tend to zero at the safe set boundary [4].

Definition 3 (*Reciprocal Control Barrier Functions (RCBFs)* [4])**.** *Consider the system (1) and the set $\mathcal{C} \subset \mathbb{R}^n$. A continuously differentiable function $B : \mathrm{Int}(\mathcal{C}) \to \mathbb{R}$ is called an RCBF if there exist a class $\mathcal{K}$ functions $\alpha_1, \alpha_2, \alpha_3$ such that, for all $x \in \mathrm{Int}(\mathcal{C})$:*

$$\frac{1}{\alpha_1(h(x))} \leq B(x) \leq \frac{1}{\alpha_2(h(x))} \tag{11}$$

$$\inf_{u \in \mathcal{U}} \left[L_f B(x) + L_g B(x) u - \alpha_3(h(x))\right] < 0 \tag{12}$$

with $L_f B(x) = \frac{\partial B(x)}{\partial (x)} f(x)$ and $L_g B(x) = \frac{\partial B(x)}{\partial (x)} g(x)$.

The RCBF B is said to be Lipschitz continuous locally if α_3 and $\frac{\partial B}{\partial x}$ are both locally Lipschitz continuous [4].

Definition 4 (Zeroing Control Barrier Functions (ZCBFs)[4])**.** *Given a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ and a set $\mathcal{C} \subset \mathbb{R}^n$ defined in* (8)*, the function h is called a ZCBF (Zeroing Control Barrier Function) defined on a larger compact set $\mathcal{D}$ with $\mathcal{C} \subseteq \mathcal{D} \subset \mathbb{R}^n$ if there exists an extended class $\mathcal{K}_\infty^{\mathrm{e}}$ function α_4 such that*

$$\sup_{u \in \mathcal{U}} \left[L_f h(x) + L_g h(x) u\right] \geq -\alpha_4(h(x)), \forall x \in \mathcal{D}. \tag{13}$$

In this case, the ZCBF h is said to be locally Lipschitz continuous if both α_4 and the derivative of h are locally Lipschitz continuous.

Control Lyapunov functions (CLF) are typically used to characterize the family of controllers that stabilize the system in the sense of Lyapunov [30].

Definition 5 (Control Lyapunov Functions (CLFs))**.** *A continuously differentiable positive definite function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a CLF for* (1) *(on Ω) if there exists $\alpha_5 \in \mathcal{K}$ such that $\forall x \in \Omega \setminus \{0\}$*

$$\inf_{u \in \mathcal{U}} \left[L_f V(x) + L_g V(x) u\right] < -\alpha_5(\|x\|) \tag{14}$$

with $L_f V(x) = \frac{\partial V(x)}{\partial (x)} f(x)$ and $L_g V(x) = \frac{\partial V(x)}{\partial (x)} g(x)$.

C. System under exploration

Consider the system (1) subjected to probing noise $e_u(t)$ during the exploration phase $\forall t \geq 0$ as:

$$\dot{x} = f(x) + g(x)(u(t) + e_u(t)) \tag{15}$$

where $e_u : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$ is a time-varying, piecewise continuous probing noise that serves as a matched disturbance. It encompasses the effects of model uncertainty and exploration. We assume that e_u is bounded, meaning $\|e_u\|_\infty \triangleq \operatorname{ess\,sup}_{t \geq 0} \|e_u(t)\|_2 < \infty$. Our primary objective is to ensure safety during the exploration phase (data collection) in the presence of time-varying, persistently exciting bounded probing noise $e_u(t)$. To achieve this, it becomes crucial to steer the system towards safety boundaries without compromising safety constraints. $\|e_u\|_\infty = \operatorname{ess\,sup}_{t \geq 0} \|e_u(t)\| < \infty$

IV. Exploration near safety Boundaries

This section introduces a safe exploration approach that ensures the control input is safe and less conservative near the boundaries, thereby encouraging exploration [3, 15].

A. Tunable Input-to-State Safe Exploration

To guarantee safety of the system (15) with probing noise $e_u(t)$ as a matched disturbance, a larger safe set $\mathcal{C}_{\xi,\mathrm{T}} \subset \mathbb{R}^n$ is considered, parameterized by $\xi \geq 0$ such that $\mathcal{C} \subseteq \mathcal{C}_{\xi,\mathrm{T}}$. This larger set $\mathcal{C}_{\xi,\mathrm{T}}$ should remain forward invariant for all $\|e_u(t)\|$ satisfying $\|e_u(t)\|_\infty \leq \xi$ to ensure safety during the data collection phase. To that end, consider a function $h_{\xi,\mathrm{T}} : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ as:

$$h_{\xi,\mathrm{T}}(x, \xi) = h(x) + \gamma_{\mathrm{T}}(h(x), \xi) \tag{16}$$

with $\gamma_{\mathrm{T}} : \mathbb{R} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, both $h_{\xi,\mathrm{T}}$ and γ_{T} continuously differentiable with respect to its first argument and $\gamma_{\mathrm{T}}(a, \cdot) \in \mathcal{K}_\infty$ for all $a \in \mathbb{R}$. Then, a larger set $\mathcal{C}_{\xi,\mathrm{T}}$ is defined as the 0-superlevel set of the function $h_{\xi,\mathrm{T}}$ as:

$$\mathcal{C}_{\xi,\mathrm{T}} \triangleq \{x \in \mathbb{R}^n : h(x) + \gamma_{\mathrm{T}}(h(x), \xi) \geq 0\} \tag{17}$$

$$\partial \mathcal{C}_{\xi,\mathrm{T}} \triangleq \{x \in \mathbb{R}^n : h(x) + \gamma_{\mathrm{T}}(h(x), \xi) = 0\} \tag{18}$$

$$\mathrm{Int}\left(\mathcal{C}_{\xi,\mathrm{T}}\right) \triangleq \{x \in \mathbb{R}^n : h(x) + \gamma_{\mathrm{T}}(h(x), \xi) > 0\}. \tag{19}$$

It is observed that $\mathcal{C} \subset \mathcal{C}_{\xi,\mathrm{T}}$ for $\xi > 0$, implying that $\mathcal{C}_{\xi,\mathrm{T}}$ is a larger set that expands monotonically with ξ. In the absence of exploration noise, i.e., when $\xi = 0$, we recover the original set ($\mathcal{C}_{\xi,\mathrm{T}} \equiv \mathcal{C}$) as $h_{\xi,\mathrm{T}}(x, 0) = h(x)$. Consequently, the closeness of $\mathcal{C}_{\xi,\mathrm{T}}$ to $\mathcal{C}$ or the difference $h_{\xi,\mathrm{T}}(x) - h(x)$ is directly determined by the smallness of γ_T.

Definition 6 (Input-to-state Safe Exploration (ISSf-Exp))**.** *The system under investigation (15) is said to undergo Input-to-State Safe Exploration (ISSf-Exp) with respect to the set $\mathcal{C}$, if there exists a function $\gamma_{\mathrm{T}}(h(x), \xi) : \mathbb{R} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ that satisfies the following conditions: (i) $\gamma_{\mathrm{T}}(a, \cdot) \in \mathcal{K}_\infty$ for all $a \in \mathbb{R}$, (ii) $\gamma_{\mathrm{T}}(\cdot, b)$ is continuously differentiable for all $b \in \mathbb{R}_{\geq 0}$; and (iii) For all $\xi \geq 0$ and $e_u(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$ satisfying $\|e_u(t)\|_\infty \leq \xi$, the set $\mathcal{C}_{\xi,\mathrm{T}}$ defined by (17)-(19) is forward invariant.*

If the system (15) is ISSf-Exp with respect to the set $\mathcal{C}$ (i.e., the set $\mathcal{C}_{\xi,\mathrm{T}}$ is forward invariant under exploration), then the set $\mathcal{C}$ is referred to as an ISSf-Exp set. Inspired by the construction of ISSf-CBF in [15] and the Tunable ISSf-CBF in [2], this paper proposes an ISSf-Exp CBF to characterize control inputs that guarantee ISSf-Exp.

Definition 7 (ISSf-Exp-CBF)**.** *Let $\mathcal{C} \subset \mathbb{R}^n$ be defined as the 0-superlevel set of a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$, where $\frac{\partial h}{\partial x}(x) \neq 0$ whenever $h(x) = 0$. The function h is an Input-to-State Safe Exploration Control Barrier Function (ISSf-Exp-CBF) for the system under exploration*

$\dot{x} = f(x) + g(x)\big(u + e_u(t)\big)$ *with* $\|e_u\|_\infty \le \xi$ *(15) over the set* $\mathcal{C}$ *with a continuously differentiable function* $\epsilon : \mathbb{R} \to \mathbb{R}_{>0}$ *if there exists* $\alpha_4 \in \mathcal{K}_\infty^{\mathrm{e}}$ *such that* $\forall x \in \Omega \in \mathbb{R}^n$

$$\sup_{u \in \mathcal{U}} \Big[L_f h(x) + L_g h(x)\, u \Big] \ge -\alpha_4\big(h(x)\big) + \frac{\|L_g h(x)\|_2^2}{\epsilon\big(h(x)\big)}. \quad (20)$$

The ISSf-Exp-CBF characterizes the set of point-wise control inputs that enable safe exploration in the sense of ISSf-Exp. It is observed that a small value of ϵ increases the right-hand side of (20), resulting in the enforcement of more stringent conditions that the control input must satisfy for safety during exploration, and vice versa.

Remark 1. It is noted that proposed ISSf-Exp-CBF specializes the ISSf-CBFs and Tunable ISSf-CBFs [2, 3], to safe RL exploration by explicitly treating the probing signal as a matched disturbance and by introducing a state-dependent tuning $\epsilon(h)$ that tightens the constraint near ∂C while relaxing it deeper in $\mathrm{Int}(C)$, thereby encouraging informative boundary-aware data collection under persistent excitation.

Next, relationship between $\gamma(\cdot)$ and α_4 is established.

Lemma 1. *Consider the system under exploration given by (15) with* $\|e_u(t)\|_\infty \le \xi$*, and* $\mathcal{C}_{\xi,\mathrm{T}}$ *being the* 0*-superlevel set of* $h_{\xi,\mathrm{T}}$ *(17). Then, given the control inputs that satisfy the ISSf-Exp CBF condition (20)* $\forall x \in \Omega \in \mathbb{R}^n$ *with* $\epsilon : \mathbb{R} \to \mathbb{R}_{>0}$ *being a continuously differentiable function and* $\alpha_4 \in \mathcal{K}_\infty^{\mathrm{e}}$*, one has*

$$\gamma_{\mathrm{T}}(h(x), \xi) \triangleq -\alpha_4^{-1}\left(-\frac{\epsilon(h(x))\xi^2}{4}\right). \quad (21)$$

Proof. Along the exploration dynamics (15), the safety function satisfies

$$\dot{h}(x,t) = L_f h(x) + L_g h(x)\, u(t) + L_g h(x)\, e_u(t).$$

Assume that the applied control input $u(t)$ satisfies the ISSf-Exp-CBF condition (20). Then

$$\begin{aligned}\dot{h}(x,t) &\ge -\alpha_4\big(h(x)\big) + \frac{\|L_g h(x)\|_2^2}{\epsilon(h(x))} + L_g h(x)\, e_u(t) \\ &\ge -\alpha_4\big(h(x)\big) + \frac{\|L_g h(x)\|_2^2}{\epsilon(h(x))} - \|L_g h(x)\|_2 \, \|e_u(t)\|_2. \end{aligned} \quad (22)$$

Using $\|e_u\|_\infty \le \xi$ and the inequality $a \ge 0,\ \epsilon > 0$,

$$\frac{a^2}{\epsilon} - a\,\xi = \left(\frac{a}{\sqrt{\epsilon}} - \frac{\sqrt{\epsilon}\,\xi}{2}\right)^2 - \frac{\epsilon\,\xi^2}{4} \ \ge\ -\frac{\epsilon\,\xi^2}{4},$$

with $a = \|L_g h(x)\|_2$ and $\epsilon = \epsilon(h(x))$, we obtain from (22):

$$\dot{h}(x,t) \ge -\alpha_4\big(h(x)\big) - \frac{\epsilon\big(h(x)\big)\,\xi^2}{4}. \quad (23)$$

Define

$$\gamma_{\mathrm{T}}\big(h(x), \xi\big) \triangleq -\alpha_4^{-1}\left(-\frac{\epsilon\big(h(x)\big)\,\xi^2}{4}\right),$$

which is nonnegative since $\alpha_4 \in \mathcal{K}_\infty^{\mathrm{e}}$ is strictly increasing and $\epsilon(h)\xi^2/4 \ge 0$. □

Remark 2. The set gap $\mathcal{C}_{\xi,\mathrm{T}} \setminus \mathcal{C}$ (or, $h_{\xi,\mathrm{T}}(x) - h(x)$) is governed by γ_{T}, which scales with the exploration noise ξ and the tuning parameter ϵ. For fixed ξ, smaller ϵ reduces γ_{T}, so $\mathcal{C}_{\xi,\mathrm{T}} \to \mathcal{C}$ and the constraint in (20) tightens, enforcing more conservative control actions, appropriate near the safety boundary. Larger ϵ increases γ_{T}, enlarging $\mathcal{C}_{\xi,\mathrm{T}}$ and relaxing feasibility, thereby allowing less conservative inputs when the state lies in the interior or moderately close to the boundary.

In this context, ϵ regulates how the noisy behavior policy is admitted during exploration. From (21) (with $\alpha_4 \in \mathcal{K}_\infty^{\mathrm{e}}$), decreasing ϵ decreases γ_{T}; hence, by (16), $\mathcal{C}_{\xi,\mathrm{T}}$ contracts toward $\mathcal{C}$, tightening feasibility for $u+e_u(t)$ in (15). Equivalently, (20) becomes more restrictive as ϵ decreases.

B. ϵ-Tuning Law for Exploration near Boundaries

An ϵ-tuning law is proposed that imposes stronger conditions on the structure of $h_{\xi,\mathrm{T}}$ and $\mathcal{C}_{\xi,\mathrm{T}}$ when the system's state is only close to the safety limit, i.e., $h(x) \equiv 0$. This leads to high conservativeness in the choice of the control input (behavior policy). In contrast, when the system states are within the safe set, that is, $h(x) > 0$, conditions are relaxed through a large value of ϵ, resulting in the desired non-conservativeness in the choice of the behavioral control policy. It is proposed to vary $\epsilon(h(x))$ as:

$$\epsilon(h) = \epsilon_0 + (\epsilon_{\max} - \epsilon_0)\left(1 - \exp\left(-\frac{\sigma(h)}{\tau}\right)\right) \quad (24)$$

$$\text{with} \quad \sigma(h) \triangleq \begin{cases} 0, & h \le 0, \\ h\,\exp\left(-\dfrac{1}{\kappa h}\right), & h > 0, \end{cases} \quad (25)$$

where $\epsilon_0 > 0$ is the minimum value attained at and outside the boundary ($h \le 0$), $\epsilon_{\max} > \epsilon_0$ is the maximum interior value, and $\tau > 0$ sets the transition depth. The function $\sigma(h)$ is a C^∞ smooth positive-part surrogate: $\sigma(h) = 0$ for $h \le 0$ and $\sigma(h) \sim h$ for large $h > 0$. We have $\kappa > 0$ that signifies a damping coefficient and in this work necessarily choose $\kappa = 10^5$ as very high value so that $\sigma(h) \to h$ rapidly with small increase in value of h. Thus $\epsilon(h) \in [\epsilon_0, \epsilon_{\max})$ for all $h \in \mathbb{R}$, with $\epsilon(h) \to \epsilon_{\max}$ as $h \to +\infty$. In particular, small values near ∂C enforce stringent exploration constraints leading to high conservatism, while larger values in $\mathrm{Int}(C)$ reduce conservatism and permit richer exploration. Next, the forward invariance of the set $\mathcal{C}_{\xi,\mathrm{T}}$ is guaranteed under the ϵ-tuning law to establish the safety of the system (15) under exploration.

Theorem 1 (ISSf Exploration)**.** *Let* $\mathcal{C} \subset \mathbb{R}^n$ *be defined as the* 0*-superlevel set of a continuously differentiable function* $h : \mathbb{R}^n \to \mathbb{R}$*, where* $\frac{\partial h}{\partial x}(x) \ne 0$ *whenever* $h(x) = 0$*, as in (8). Assume further that* 0 *is a regular value of* $h_{\xi,\mathrm{T}}(\cdot, \xi)$*, i.e.,* $\frac{\partial h_{\xi,\mathrm{T}}}{\partial x}(x, \xi) \ne 0$ *for all* $x \in \partial\mathcal{C}_{\xi,\mathrm{T}}$*. Suppose that* h *is an ISSf-Exp-CBF (see Definition 7) for the system (15) over the set* $\mathcal{C}$ *with a continuously differentiable function* $\epsilon : \mathbb{R} \to \mathbb{R}_{>0}$ *defined as in (24) with* ϵ_0*,* ϵ_{max}*,* τ *as positive constants, then* $\forall x \in \Omega \subseteq \mathbb{R}^n$ *and for all* $\|e_u\|_\infty \le \xi$ *the system (15) undergoes Input-to-State Safe Exploration (ISSf-Exp) with respect to the set* $\mathcal{C}$ *i.e.,* $\mathcal{C}_{\xi,\mathrm{T}}$ *is forward invariant under exploration with* $\gamma_{\mathrm{T}} : \mathbb{R} \times \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ *defined in (21).*

Proof. Taking the time derivative of $h_{\xi,\mathrm{T}}$ yields:

$$\dot{h}_{\xi,\mathrm{T}}(x,\xi,t) = \left(1 + \frac{\partial \gamma_{\mathrm{T}}}{\partial h}(h(x),\xi)\right)\dot{h}(x,t). \tag{26}$$

From the expression of $\gamma_{\mathrm{T}}(h(x),\xi)$ in (21) and (24), it is clear that

$$(1 + \frac{\partial \gamma_{\mathrm{T}}}{\partial h}(h(x),\xi)) > 0. \tag{27}$$

On $\partial\mathcal{C}_{\xi,\mathrm{T}}$ we have $h(x) + \gamma_{\mathrm{T}}(h(x),\xi) = 0$, i.e., $h(x) = -\gamma_{\mathrm{T}}(h(x),\xi)$. Using the definition (21), this boundary condition implies

$$h(x) = \alpha_4^{-1}\left(-\frac{\epsilon(h(x))\,\xi^2}{4}\right) \tag{28}$$

which implies

$$\alpha_4(h(x)) = -\frac{\epsilon(h(x))\,\xi^2}{4}. \tag{29}$$

Hence, from (23) we obtain on $\partial\mathcal{C}_{\xi,\mathrm{T}}$:

$$\dot{h}(x,t) \geq -\alpha_4(h(x)) - \frac{\epsilon(h(x))\,\xi^2}{4} = 0.$$

Since (27) ensures $1 + \frac{\partial \gamma_{\mathrm{T}}}{\partial h} > 0$, it follows from (26) that, on $\partial\mathcal{C}_{\xi,\mathrm{T}}$,

$$\dot{h}_{\xi,\mathrm{T}}(x,\xi,t) \geq 0. \tag{30}$$

Further, consider

$$\frac{\partial h_{\xi,\mathrm{T}}}{\partial x}(x,\xi) = \left(1 + \frac{\partial \gamma_{\mathrm{T}}}{\partial h}(h(x),\xi)\right)\frac{\partial h}{\partial x}(x). \tag{31}$$

We have, by assumption, $\frac{\partial h_{\xi,\mathrm{T}}}{\partial x}(x,\xi) \neq 0$. Thus, $h_{\xi,\mathrm{T}} = 0$ implies $\dot{h}_{\xi,\mathrm{T}}(x,\xi,t) \geq 0$ and $\frac{\partial \dot{h}_{\xi,\mathrm{T}}}{\partial x}(x,\xi) \neq 0$. Then, after using Nagumo's theorem [21, 24], the forward invariance of the set $\mathcal{C}_{\xi,\mathrm{T}}$ is guaranteed. □

C. Robust QP based Safe Exploration

Let $u_0 : \Omega \to \mathbb{R}^m$ denote a baseline (behavior) feedback policy. During data collection, we inject a bounded probing signal $e_u(t)$ and define the pre-filter input

$$u_{\mathrm{noisy}}(t) \triangleq u_0(x(t)) + e_u(t).$$

To enforce stability of the equilibrium point of the closed-loop system as well as tunable safety, we will synthesize the control law during exploration by solving a QP optimization [4, 8, 9] that unifies the stability and safety objectives through CLFs and CBFs respectively. The proposed QP optimization is given as:

$$\begin{aligned}
&(u_{QP}^{\star},\, \delta^{\star}) = \arg\min_{u_{QP}\in\mathbb{R}^m,\, \delta\in\mathbb{R}} \frac{1}{2}\|u_{QP}\|_2^2 + \frac{p}{2}\,\delta^2 \\
&\text{s.t.} \\
&L_f V(x) + L_g V(x)\big(u_{\text{noisy}}(t) + u_{QP}\big) \leq -\,\alpha_5\big(\|x\|\big) + \delta, \\
&L_f h(x) + L_g h(x)\big(u_{\text{noisy}}(t) + u_{QP}\big) \\
&\geq -\,\alpha_4\big(h(x)\big) + \frac{\|L_g h(x)\|_2^2}{\epsilon\big(h(x)\big)}, \\
&-\,\rho\,\mathbf{1} \leq u_{\text{noisy}}(t) + u_{QP} \leq \rho\,\mathbf{1},
\end{aligned} \tag{32}$$

where $u_{\mathrm{noisy}}(t) \triangleq u_0(x(t)) + e_u(t)$ is the known pre-filter noisy control input prior to QP-based filtering. The applied input is $u_{\text{applied}}(t) \triangleq u_{\mathrm{noisy}}(t) + u_{QP}(t)$, where $u_{QP}(t) \triangleq u_{QP}^{\star}(x(t), u_{\mathrm{noisy}}(t))$ is the optimizer of (32).
Moreover, $\alpha_4 \in \mathcal{K}_{\infty}^{\mathrm{e}}, \alpha_5 \in \mathcal{K}$, δ is a relaxation variable that softens the CLF constraint to preserve feasibility while keeping the safety (ISSf-Exp-CBF) constraint hard; $p \gg 1$ penalizes CLF violation [9]. The constraints consist of CLF (cf. (14)) and ISSf-Exp-CBF conditions (cf. (20)) to enforce stability and safety respectively.

Remark 3. Solving (32) point-wise yields the QP correction $u_{QP}(t) \triangleq u_{QP}^{\star}(x(t), u_{\mathrm{noisy}}(t))$, and the actual plant input is $u_{\mathrm{applied}}(t) \triangleq u_{\mathrm{noisy}}(t) + u_{QP}(t)$. To ensure strict compliance with input saturation, the saturation bounds are encoded as hard constraints within the ISSF-Exp-QP problem. The QP in (32) acts as an online safety filter, where $u_{QP}^{\star}$ is the minimum-norm correction that enforces the CLF and ISSf-Exp-CBF inequalities point-wise.

Remark 4. For initializing policy iteration (which requires a state-feedback policy), we define the filtered baseline

$$u^{(0)}(x) \triangleq u_0(x) + u_{QP}^{\star}(x, u_0(x)),$$

i.e., the same QP filter applied with $e_u \equiv 0$.

Lemma 2. *Given the system under exploration (15), the ISSf-Exp-QP problem in (32) yields an optimizer mapping $(x, u_{\mathrm{noisy}}) \mapsto u_{QP}^{\star}(x, u_{\mathrm{noisy}})$ that is locally Lipschitz under standard parametric QP regularity conditions. The resulting applied input $u_{\mathrm{applied}}(t) = u_{\mathrm{noisy}}(t) + u_{QP}^{\star}(x(t), u_{\mathrm{noisy}}(t))$ renders the enlarged set $\mathcal{C}_{\xi,\mathrm{T}}$ forward invariant and guarantees ISSf-Exp with respect to $\mathcal{C}$ in the sense of Definition 6.*

Proof. The constraints in (32) are linearly independent; under standard parametric QP regularity conditions, the optimizer mapping $(x, u_{\mathrm{noisy}}) \mapsto u_{\mathrm{QP}}^{\star}(x, u_{\mathrm{noisy}})$ is locally Lipschitz. For sufficiently large design parameter p, $u_{\mathrm{QP}}^{\star}$ approximates the min-norm controller of [7] and is therefore asymptotically stabilizing [9]. Moreover, if there exists any control satisfying (20), then the relaxation variable δ ensures feasibility of the CLF constraint in (32), implying the existence of at least one QP correction $u_{QP}^{\star}$ satisfying both the CLF and ISSf-Exp-CBF constraints. Consequently, the ISSf-Exp-CBF condition renders $\mathcal{C}_{\xi,\mathrm{T}}$ forward invariant and guarantees that (15) is ISSf-Exp with respect to $\mathcal{C}$. □

V. Safe Learning

Consider a safety-aware utility function r_{safe} that has been augmented by an RCBF candidate function as:

$$r_{\text{safe}}(x,u) = Q(x) + B_\lambda(h(x)) + 2\int_0^u \left(\rho\tanh^{-1}(\frac{v}{\rho})\right)^{\top} R\mathrm{d}v \tag{33}$$

wherein $B_\lambda(\cdot)$ is an RCBF candidate (cf. (11)). In this work, we use [12]:

$$B_\lambda(h(x)) = -\log\left(\frac{\lambda h(x)}{\lambda h(x) + 1}\right). \tag{34}$$

where $\lambda > 0$ is a damping coefficient and affects how rapidly BF varies with $h(x)$ (cf. [10, 12]. Then, safety of systems

under input saturation during the learning stage can be guaranteed by considering a modified safety-aware performance index as:

$$V_{\text{safe}}(x,u) = \int_t^\infty \Bigg(Q(x(\tau)) + B_\lambda(h(x(\tau))) + 2\int_0^{u(x(\tau))} \left(\rho \tanh^{-1}(\frac{v}{\rho})\right)^\top R \mathrm{d}v \Bigg) \mathrm{d}\tau. \tag{35}$$

The reward in (2) is augmented with an RCBF candidate term. While the state remains well within the safe set, this term stays near zero; as the trajectory approaches the safety boundary, it grows, thereby sharply increasing the objective. Consequently, minimizing (35) leads to learning of policies that keep the state in the interior of the safe set [12, 19].

Remark 5 (Exploration set vs. learning objective)**.** The enlarged set $C_{\xi,T}$ is used only during data collection to guarantee robust invariance under the bounded probing signal in the exploration dynamics. In contrast, the RCBF augmentation in r_{safe} and V_{safe} is defined with respect to the original specification C to ensure that the learned policy is optimized against the true safety constraint, rather than against the robustness enlargement introduced for exploration.

Definition 8 (Safe inputs)**.** *A feedback policy $u(\cdot)$ is called safe on Ω if for every $x_0 \in \mathrm{Int}(\mathcal{C}) \cap \Omega$, the corresponding closed-loop solution satisfies*

$$x(t; x_0, u) \in \mathrm{Int}(\mathcal{C}), \qquad \forall t \geq 0.$$

Denote by $\pi_s(\Omega)$ the set of all such safe policies.

Definition 9 (Admissible Safe inputs)**.** *Admissible control policy for the safety-aware cost function under saturation (35) is defined as the set of control inputs $\pi_a(\Omega)$ that is admissible (1) as well as safe (8), i.e, $\pi_a(\Omega) = \{u \in \pi(\Omega) \cap \pi_s(\Omega) \mid V_{\text{safe}}(x,u) < \infty,\ \forall x \in \Omega\}$.*

Assumption 2. There exists at least one admissible safe feedback policy on Ω, i.e., $\pi_a(\Omega) = \pi(\Omega) \cap \pi_s(\Omega) \neq \varnothing$.

First, note that the QP filter (32) provides a safe applied input during data collection via $u_{\text{applied}}(t) = u_{\text{noisy}}(t) + u_{QP}(t)$. For the learning stage (which requires a state-feedback initial policy), we define an initial admissible safe policy $u^{(0)}$ by filtering a baseline feedback through the same QP map, as stated next.

Lemma 3. *Let $u_0 : \Omega \to \mathbb{R}^m$ be a continuous baseline (behavior) feedback and let*

$$u^{(0)}(x) \triangleq u_0(x) + u^\star_{QP}(x, u_0(x)),$$

where $u^\star_{QP}(\cdot,\cdot)$ is obtained by solving the ISSf-Exp-QP (32) *point-wise (with $u_{\text{noisy}} = u_0$). Assume $x_0 \in \mathrm{Int}(\mathcal{C}) \cap \Omega$. Then $u^{(0)} \in \pi_a(\Omega)$, and $V_{\text{safe}}(x, u^{(0)})$ in* (35) *is finite on Ω i.e. $u^{(0)}$ is admissible and safe for the safety-aware cost function under saturation (35).*

Proof. The proof directly follows from Lemma 2. □

Next, the existence of a value function with respect to the safety problem (2) is established.

Lemma 4. *Under control input saturation, given an initial safe and admissible control policy $u^{(0)} \in \pi_a(\Omega)$ (e.g., as constructed in Lemma 3), if*

$$\begin{aligned} &\nabla W^\top(x)(f(x) + g(x)u^{(0)}) + Q(x) + B_\lambda(h(x)) \\ &+ 2\int_0^{u^{(0)}} \left(\rho \tanh^{-1}(\frac{v}{\rho})\right)^\top R \mathrm{d}v = 0 \\ &W(0) = 0. \end{aligned} \tag{36}$$

then, W is the value function of the system (1) under input saturation for all $t \geq 0$, then $W(x) = V_{\text{safe}}(x, u^{(0)})$ on Ω.

Proof. Assume that $W\left(x, u^{(0)}\right) > 0$ such that $W \in C^1$, then integrating the time derivative along the flow of the system under the control policy $u^{(0)}$, one has

$$\begin{aligned} W\left(x(t), u^{(0)}\right) - W\left(x_0, u^{(0)}\right) &= \int_0^T \dot{W}\left(x(\tau), u^{(0)}\right) \mathrm{d}\tau \\ &= \int_0^T \frac{\partial W}{\partial x}\left(f + g u^{(0)}\right) \mathrm{d}\tau. \end{aligned} \tag{37}$$

Further, considering (35), one has

$$V_{\text{safe}}\left(x(t), u^{(0)}\right) - V_{\text{safe}}\left(x_0, u^{(0)}\right) = -\int_0^T r_{\text{safe}}(x(\tau), u^{(0)}) \mathrm{d}\tau. \tag{38}$$

Subtracting both sides of (38) from (37) yields

$$\begin{aligned} &V_{\text{safe}}\left(x(t), u^{(0)}\right) - W\left(x(t), u^{(0)}\right) \\ &= \int_0^T \left(-\frac{\partial W}{\partial x}\left(f + g u^{(0)}\right) - r_{\text{safe}}\left(x(\tau), u^{(0)}\right)\right) \mathrm{d}\tau \\ &+ V_{\text{safe}}\left(x_0, u^{(0)}\right) - W\left(x_0, u^{(0)}\right). \end{aligned} \tag{39}$$

Considering (33) and (36) in (39) gives

$$\begin{aligned} &V_{\text{safe}}\left(x, u^{(0)}\right) - W\left(x, u^{(0)}\right) \\ &= \int_0^T r_{\text{safe}}\left(x(\tau), u^{(0)}\right) - r_{\text{safe}}\left(x(\tau), u^{(0)}\right) \mathrm{d}\tau = 0 \end{aligned} \tag{40}$$

yielding

$$V_{\text{safe}}\left(x, u^{(0)}\right) = W\left(x, u^{(0)}\right).$$

Therefore, for the fixed admissible safe policy $u^{(0)}$, the function $W(x)$ coincides with the associated value function, i.e., $W(x) = V_{\text{safe}}(x, u^{(0)})$ on Ω. □

Now, that the existence of positive definite value function has been established, a Generalized Safety-aware HJB under saturation (G-SHJB) is proposed.

Definition 10 (Generalized Safety-aware HJB)**.** *A Generalized Safety-aware HJB (G-SHJB) under input saturation is defined taking into account the positive definite value function (36) as:*

$$\begin{aligned} &\nabla W^T(x)(f(x) + g(x)u) + Q(x) + B_\lambda(h(x)) \\ &+ 2\int_0^{u} \left(\rho \tanh^{-1}(\frac{v}{\rho})\right)^\top R \mathrm{d}v = 0 \\ &W(0) = 0 \end{aligned} \tag{41}$$

with $\nabla W(x) = \partial W(x)/\partial x \in \mathbb{R}^n$.

Remark 6. The G-SHJB with boundary conditions provides a characterization of infinite-horizon optimal control. For any

admissible policy, solving (41) yields a positive definite value function $W(x)$; however, closed-form solutions are generally unavailable, motivating iterative schemes such as successive approximation [1, 16]. Prior GHJB-based successive-approximation methods addressed constrained optimal control without explicit safety guarantees [1, 29]. Here, we adopt a similar successive-approximation framework, to address safe learning problem under saturation.

The Safe Policy Iteration (S-PI) framework is shown in Algorithm 1.

Algorithm 1 Safe Policy Iteration with Input Saturation

Require:

1: Set the iteration index $i = 0$ and start with admissible initial policy $u^{(0)}(\cdot)$;

2: *Policy Evaluation*: Given the control input $u^{(i)} \in \pi_a(\Omega)$, find the value function $W^{(i)}$ by solving G-SHJB $(W^{(i)}, u^{(i)}) = 0$ i.e.,

$$\begin{aligned} &\nabla W^{(i)T}(x)(f(x)+g(x)u^{(i)}) + Q(x) + B_\lambda(h(x)) \\ &+ 2\int_0^{u^{(i)}} \left(\rho \tanh^{-1}(\frac{v}{\rho})\right)^\top R \mathrm{d}v = 0. \\ &W^{(i)}(0) = 0 \end{aligned} \tag{42}$$

3: *Policy Improvement*: The control policy $u^{(i)}$ is improved to $u^{(i+1)}$ using:

$$u^{(i+1)}(x) = -\rho \tanh\left(\frac{1}{2\rho} R^{-1} g^\top(x) \nabla W^{(i)}(x)\right). \tag{43}$$

4: Stop if convergence is achieved; Otherwise, set $i = i + 1$ and go to Step 2;

Remark 7. It is noted that i represents an iterative step in policy evaluation and improvement. An admissible safe initial policy $u^{(0)} \in \pi_a(\Omega)$ can be obtained by filtering a baseline feedback through the QP map, as in Lemma 3, i.e., $u^{(0)}(x) = u_0(x) + u^\star_{QP}(x, u_0(x)) \in \pi_a(\Omega)$.

Next, it is shown that the sequential improvement of the control policy, as specified in equations (42) and (43), guarantees the safety, stability, and optimal performance of the system.

A. Convergence

Lemma 5. *Assume $u^{(i)} \in \pi_a(\Omega)$ and let $W^{(i)}$ solve the policy-evaluation equation* (42)*. Define the (safety-aware) Hamiltonian*

$$\begin{aligned} H\big(x, u, \nabla W^{(i)}\big) &\triangleq Q(x) + B_\lambda\big(h(x)\big) \\ &+ U(u) + \nabla W^{(i)\top}(x)\big(f(x) + g(x)u\big) \end{aligned} \tag{44}$$

with U(u) defined in (3)*. Let $u^{(i+1)}$ be given by the policy improvement step* (43)*. Then:*

1) *$u^{(i+1)}$ is continuous and satisfies the saturation constraint $u^{(i+1)}(x) \in \mathcal{U}$ for all $x \in \Omega$.*
2) *$u^{(i+1)} \in \pi_a(\Omega)$, and the corresponding value function satisfies $W^{(i+1)}(x) \le W^{(i)}(x)$ for all $x \in \Omega$.*
3) *If $x_0 \in \mathrm{Int}(\mathcal{C})$, then $B_\lambda\big(h(x(t))\big)$ remains finite along the closed-loop trajectory under each $u^{(i)}$.*

Proof. By construction, (43) is the point-wise minimizer of $u \mapsto H(x, u, \nabla W^{(i)})$ under the saturated integrand $U(u)$, hence

$$H\big(x, u^{(i+1)}, \nabla W^{(i)}\big) \le H\big(x, u^{(i)}, \nabla W^{(i)}\big) = 0,$$

where the last equality follows from (42). Along trajectories of $\dot{x} = f(x) + g(x)u^{(i+1)}(x)$,

$$\begin{aligned} &\frac{d}{dt} W^{(i)}(x(t)) = \nabla W^{(i)\top}(x)\big(f(x) + g(x)u^{(i+1)}(x)\big) \\ &= H\big(x, u^{(i+1)}, \nabla W^{(i)}\big) - Q(x) - B_\lambda(h(x)) - U(u^{(i+1)}), \end{aligned} \tag{45}$$

and therefore $\dot{W}^{(i)}(x(t)) \le -Q(x(t)) - B_\lambda(h(x(t))) - U(u^{(i+1)}(x(t))) \le 0$. This implies stability and admissibility of $u^{(i+1)}$ on Ω. Integrating the above inequality from 0 to ∞ yields

$$\begin{aligned} W^{(i)}(x_0) &\ge \int_0^\infty \big(Q(x(t)) + B_\lambda(h(x(t))) + U(u^{(i+1)}(x(t)))\big)\, dt \\ &= W^{(i+1)}(x_0). \end{aligned} \tag{46}$$

hence $W^{(i+1)} \le W^{(i)}$ on Ω. Finally, since $W^{(i)}(x_0) < \infty$ for $x_0 \in \mathrm{Int}(\mathcal{C})$ and $B_\lambda(h) \to \infty$ as $h \to 0^+$ for a reciprocal barrier candidate, the trajectory cannot reach $\partial\mathcal{C}$ in finite time; thus $B_\lambda(h(x(t)))$ remains finite and $x(t) \in \mathrm{Int}(\mathcal{C})$. □

Theorem 2 (Safety)**.** *Assume $x_0 \in \mathrm{Int}(\mathcal{C})$ and that $B_\lambda(\cdot)$ is a reciprocal barrier candidate satisfying $B_\lambda(h) \to \infty$ as $h \to 0^+$. Then, under the policy sequence generated by* (42)*-*(43)*, the closed-loop trajectory satisfies $x(t) \in \mathrm{Int}(\mathcal{C})$ for all $t \ge 0$.*

Proof. Lemma 5 demonstrates that the value function $W^{(i)}$ and the RCBF candidate function $B_\lambda^{(i)}$ remain bounded at each iterative step i during sequential improvement steps (42) and (43). Conversely, given the properties of the RCBF, the value of B_λ^i approaches infinity when the system states approach the boundary. However, since the RCBF remains bounded at each iterative step i, it is guaranteed that the states do not reach the boundary. Consequently, the safety of the system is guaranteed. □

B. Stability

Definition 11. *Feasible region [18] Define the interior of the safe set $\mathrm{Int}(\mathcal{C})$ (8) as a feasible set.*

Definition 12 (Safe region [18])**.** *Let $\beta(\partial\mathcal{C}, r_0) \triangleq \{x \in \mathbb{R}^n \mid \mathrm{dist}(x, \partial\mathcal{C}) < r_0\}$ denote the open r_0-neighborhood of the boundary. Define the safe region as*

$$D \triangleq \mathrm{Int}(\mathcal{C}) \setminus \beta(\partial\mathcal{C}, r_0),$$

and assume $0 \in D$.

Remark 8. The damping factor is chosen in a manner that ensures $Q(x)$ dominates $B_\lambda(h(x))$ within the safe region, as discussed in [18]. Furthermore, it is assumed that safety does not compromise performance within the safe region. The trade-off between $Q(x)$ and $B_\lambda(h(x))$ within the safe region is determined by the coefficient λ. Higher values of λ accelerate the damping of $B_\lambda(h(x))$ as it moves further from

the boundary, while preserving the original safety-agnostic utility function $r(x,u) = Q(x) + 2\int_0^u \left(\rho \tanh^{-1}(\frac{v}{\rho})\right)^\top R\mathrm{d}v$. Conversely, smaller values of λ place greater emphasis on safety, leading to a more conservative control design.

Theorem 3 (Stability)**.** *Assume that $x = 0$ is the equilibrium point of the system (1), and $D \subset \Omega$ contains the origin. Then, given the policy sequence obtained using the Safe PI algorithm (43), $\{u^{(i)}\}_{i=1}^{i+1} = \{u^{(1)}, u^{(2)} ... u^{(i+1)}\}$ and the corresponding sequence of positive value functions be $\{W^{(i)}\}_{i=1}^{i=i+1} = \{W^{(1)}, W^{(2)} ... W^{(i+1)}\}$, the system is uniformly stable within the safe region D.*

Proof. From Lemma 5 (first part), we have that $\dot{W}^{(i)}\left(x, u^{(i)}\right) = \nabla W^{(i)\top}(x)(f + gu^{(i)}) \leq 0$. Moreover, from Lemma 5 (second part), it is shown that $W^{(i)}(x)$ is a Lyapunov function for $u^{(i+1)}$ on Ω or, $W^{(i+1)} \leq W^{(i)}\ \forall x \in \Omega$. Then, from the definition of safe region D (Definition 12) and from Theorem 2, one has $W^{(i+1)} \leq W^{(i)}\ \forall x \in D$. Now, since $W^{(i)}$ is bounded (see Theorem 2), since $W^{(1)}$ is continuous and positive definite on the compact set D, there exist class-$\mathcal{K}$ functions $\underline{\alpha}, \overline{\alpha}$ such that $\underline{\alpha}(\|x\|) \leq W^{(1)}(x) \leq \overline{\alpha}(\|x\|)$ for all $x \in D$. Thus, it can be shown in the sense of proof of Theorem 4.8 on p. 151 of [13], that the origin is uniformly stable. □

C. Optimality

An optimal control law $u^\star(x)$ and the corresponding optimal value function $W^\star(x)$ must adhere to (36). To achieve this, a Safety-aware Hamiltonian-Jacobi Equation (SHJB) is defined as follows.

Definition 13. *The SHJB associated with optimal control law $u^\star(x)$ and $W^\star(x)$ is defined as :*

$$\begin{aligned} &\nabla W^{\star\top}(x)(f(x) + g(x)u^\star) + Q(x) + B_\lambda(h(x)) \\ &+ 2\int_0^{u^\star} \left(\rho \tanh^{-1}(\frac{v}{\rho})\right)^\top R\mathrm{d}v = 0 \\ &W^\star(0) = 0 \end{aligned} \tag{47}$$

with the optimal control law $u^\star(x)$ given as:

$$u^\star(x) = -\rho \tanh\left(\frac{1}{2\rho} R^{-1} g^\top(x) \nabla W^\star(x)\right). \tag{48}$$

Remark 9. The G-SHJB is solved iteratively using (42) and (43) until the optimal value function $W^\star$ is obtained. Subsequently, it is demonstrated that if the initial control law is both safe and admissible, then the iterative procedure leads to uniqueness and optimality, i.e., $W^{(i)} \to W^\star$.

Theorem 4 (Optimality)**.** *Suppose that the equilibrium point of the system (1) is $x = 0$, and $D \subset \Omega$ contains the origin. Given the policy sequence obtained using the safe PI algorithm (43), we have $\{u^{(i)}\}_{i=1}^{i+1} = \{u^{(1)}, u^{(2)} \dots u^{(i+1)}\}$ and the corresponding sequence of positive value functions $\{W^{(i)}\}_{i=1}^{i=i+1} = \{W^{(1)}, W^{(2)} \dots W^{(i+1)}\}$. The sequence of solutions converges to optimality, meaning that the sequence of value functions $W^{(i)}$ and the sequence of control laws $u^{(i)}$ converge, respectively, to an optimal value function $W^\star$ and a corresponding optimal safe control law $u^\star$.*

Proof. From Lemma 5, for $u^{(i)} \in \pi_a(\Omega)$ with $W^{(i)}$ being positive definite (see Lemma 4) satisfying the equation G-SHJB $\left(W^{(i)},\ u^{(i)}\right) = 0$, one has $W^{(i+1)} \leq W^{(i)}\ \forall x \in D \subseteq \Omega$. By induction, we have $W^{(i+1)} \leq W^{(i)} \leq W^{(0)}$ and $u^{(i)} \in \pi_a(\Omega)$ and through generalization, one can show $W^\star \leq W^{(i)}\ \forall i \geq 0$; leading to $W^{(i)}$ being a monotonically decreasing function that is bounded below. Then, as $i \to \infty$, the sequence will converge to some positive definite function W^∞. From [29], W^∞ is unique and must be shown to converge to the optimal solution. To this end, consider $W^{(i+1)} = W^{(i)} = W^\infty$, then from (46) one has $u^{(i+1)} = u^{(i)} = u^\infty$. Then, from (43) one has: $u^\infty(x) = -\rho \tanh\left(\frac{1}{2\rho} R^{-1} g^\top(x) \nabla W^\infty(x)\right)$. Further, G-SHJB$(W^\infty, u^{(\infty)}) = 0$ can be expressed as in (42). Clearly, those are (47) and (48) with a unique solution $W^\star$ such that $W^\star = W^\infty$ [17] and corresponding optimal safe control law $u^\star$. This, in turn implies that $W^{(i)} \to W^\star$ and $u^{(i)} \to u^\star$. □

Theorem 5. *Assume that Theorems 3-4 hold, then the minimum of the safety-aware utility function (augmented with RCBF) r_{safe} (33) can be driven arbitrarily close to zero by appropriate selection of the CBF parameter λ.*

Proof. The proof follows from [19]. □

VI. Safe Off-Policy RL

Off-policy RL is a policy-based approach that aims to find the optimal controller without requiring knowledge of the system's dynamics. It employs two policies: a behavior policy, which is safely applied to collect data (exploration), and a target policy, which is iteratively updated to optimality using the gathered data [11, 12, 19]. To that end, the nominal system dynamics (1) is rewritten to separate the behavior policy and the target policy $\forall t \geq 0$ as:

$$\dot{x} = f(x) + g(x)u^{(i)}(x(t)) + g(x)\big(u_{\text{applied}}(t) - u^{(i)}(x(t))\big), \tag{49}$$

where $u^{(i)}(x(t))$ is the target policy, which is updated in the algorithm but not applied to the system; while $u_{\text{applied}}(t)$ is the behavior input actually applied to the system (after QP filtering) to generate data for learning.

As in classical RL [12, 16], the unknown value function ($W^{(i)}$) and the control policy ($u^{(i+1)}$) are approximated using neural networks. Let Ω be a compact containing the origin as an interior point. The value function W_i (referred to as the critic) and the control policy u_{i+1} (referred to as the actor) can be approximated using the basis function representation.

$$\hat{W}^{(i)}(x) = \hat{C}_i^\top \Phi(x). \tag{50a}$$

$$\hat{u}^{(i+1)}(x) = \rho \tanh\left(\hat{U}_i \Psi(x)\right) \tag{50b}$$

with $\Phi = [\phi_1, \phi_2 \dots \phi_{N_1}]^\top$ and $\Psi = [\psi_1, \psi_2 \dots \psi_{N_2}]^\top$as two finite vectors of linearly independent smooth basis/activation functions on Ω that approximate the value function and the control policy using the Weierstrass approximation Theorem [11, 16]. The selection of $\Phi(\cdot)$ (critic) and $\Psi(\cdot)$ (actor) is a design hyperparameter, analogous to choosing network architecture in standard function approximation. To determine the weights of the critic and actor networks, a novel least squares

(LS) based closed-form formulation is developed under input saturation.

Lemma 6. *The weights of the critic ($\hat{C}_i$) and actor ($\hat{U}_i$) networks can be obtained by solving the least-squares (LS) equation*

$$\tilde{\Theta}_i^N \begin{bmatrix} \mathrm{vec}(\hat{C}_i) \\ \mathrm{vec}(\hat{U}_i^\top) \end{bmatrix} = \tilde{E}_i^N, \tag{51}$$

for $N \geq N_1 + mN_2$, provided the stacked regressor matrix $\tilde{\Theta}_i^N$ has full column rank, where

$$\begin{aligned} \tilde{\Theta}_i^N &= [\tilde{\Theta}_i(t_1), \ldots, \tilde{\Theta}_i(t_N)]^\top, \\ \tilde{E}_i^N &= [\tilde{E}_i(t_1), \ldots, \tilde{E}_i(t_N)]^\top. \end{aligned} \tag{52}$$

Here $u_{applied}(t)$ denotes the applied (behavior) input (i.e., $u_{applied}(t) = u_{\mathrm{noisy}}(t) + u_{QP}(t)$) and $v^{(i)}(t) \triangleq u_{applied}(t) - u^{(i)}(x(t))$ is the off-policy deviation (known from data). Then

$$\tilde{\Theta}_i(t) = \begin{bmatrix} \big(\Phi(x(t+T)) - \Phi(x(t))\big)^\top, \\ 2\rho \int_t^{t+T} \Big((v^{(i)}(\tau))^\top R \otimes \Psi(x(\tau))^\top \Big)\, d\tau \end{bmatrix}^\top, \tag{53}$$

and

$$\tilde{E}_i(t) = -\int_t^{t+T} \Big(Q(x(\tau)) + B_\lambda(h(x(\tau))) + U(u^{(i)}(x(\tau))) \Big)\, d\tau, \tag{54}$$

with $U(\cdot)$ given by (3)*.*

Proof. Consider the off-policy dynamics for $t \geq 0$: $\dot{x} = f(x) + g(x)u^{(i)}(x) + g(x)v^{(i)}(t)$ with $v^{(i)}(t) \triangleq u_{\mathrm{applied}}(t) - u^{(i)}(x(t))$ where $u_{\mathrm{applied}}(t)$ is the applied behavior input and $u^{(i)}$ is the target policy at iteration i. The policy-evaluation (G-SHJB) equation for $W^{(i)}$ associated with $u^{(i)}$ is

$$\begin{aligned} \nabla W^{(i)\top}(x)\big(f(x) + g(x)u^{(i)}(x)\big) = \\ -Q(x) - B_\lambda(h(x)) - U(u^{(i)}(x)). \end{aligned} \tag{55}$$

Along the off-policy trajectory,

$$\begin{aligned} \frac{d}{dt} W^{(i)}(x(t)) =& \nabla W^{(i)\top}(x(t))\dot{x}(t) \\ &= \nabla W^{(i)\top}\big(f + gu^{(i)}\big) + \nabla W^{(i)\top} g\, v^{(i)}. \end{aligned} \tag{56}$$

Using (55) yields

$$\begin{aligned} \frac{d}{dt} W^{(i)}(x(t)) = &-\Big(Q(x) + B_\lambda(h(x)) + U(u^{(i)})\Big) \\ &+ \nabla W^{(i)\top}(x) g(x) v^{(i)}(t). \end{aligned} \tag{57}$$

Integrating (57) over $[t, t+T]$ gives

$$\begin{aligned} &W^{(i)}(x(t+T)) - W^{(i)}(x(t)) = \\ &-\int_t^{t+T} \Big(Q(x) + B_\lambda(h(x)) + U(u^{(i)})\Big)\, d\tau \\ &+ \int_t^{t+T} \nabla W^{(i)\top}(x(\tau)) g(x(\tau)) v^{(i)}(\tau)\, d\tau. \end{aligned} \tag{58}$$

Since $\Gamma = \tanh$ is strictly increasing with range $(-1, 1)$, we have $|u_j^{(i+1)}(x)| < \rho$ for all x under the actor parameterization, and hence $\tanh^{-1}(u^{(i+1)}/\rho)$ is well-defined element-wise. Next, the policy improvement step (43) implies the identity

$$g^\top(x)\nabla W^{(i)}(x) = -2\rho\, R\, \tanh^{-1}\left(\frac{u^{(i+1)}(x)}{\rho}\right), \tag{59}$$

and therefore

$$\nabla W^{(i)\top}(x) g(x) = -2\rho\, \tanh^{-1}\left(\frac{u^{(i+1)}(x)}{\rho}\right)^\top R. \tag{60}$$

Substituting (60) into (58) yields

$$\begin{aligned} &W^{(i)}(x(t+T)) - W^{(i)}(x(t)) = \\ &-\int_t^{t+T} \Big(Q(x) + B_\lambda(h(x)) + U(u^{(i)})\Big)\, d\tau \\ &- 2\rho \int_t^{t+T} \tanh^{-1}\left(\frac{u^{(i+1)}(x(\tau))}{\rho}\right)^\top R\, v^{(i)}(\tau)\, d\tau. \end{aligned} \tag{61}$$

Now introduce the critic approximation $\hat{W}^{(i)}(x) = \hat{C}_i^\top \Phi(x)$ and the actor approximation $u^{(i+1)}(x) = \rho \tanh(\hat{U}_i \Psi(x))$. Since $\tanh^{-1}\left(\frac{u^{(i+1)}(x)}{\rho}\right) = \hat{U}_i \Psi(x)$, (61) becomes

$$\begin{aligned} &\hat{C}_i^\top \big(\Phi(x(t+T)) - \Phi(x(t))\big) \\ &+ 2\rho \int_t^{t+T} \Psi(x(\tau))^\top \hat{U}_i^\top R\, v^{(i)}(\tau)\, d\tau \\ &= -\int_t^{t+T} \Big(Q(x) + B_\lambda(h(x)) + U(u^{(i)})\Big)\, d\tau. \end{aligned} \tag{62}$$

Note that we parameterize the improved policy as $u^{(i+1)}(\cdot) \equiv \hat{u}^{(i+1)}(\cdot)$ with weights $\hat{U}_i$. Finally, using the identity $a^\top X b = (b^\top \otimes a^\top)\mathrm{vec}(X)$ with $a = \Psi(x(\tau))$, $X = \hat{U}_i^\top$, and $b = R\, v^{(i)}(\tau)$ yields:

$$\begin{aligned} &\Psi(x(\tau))^\top \hat{U}_i^\top R\, v^{(i)}(\tau) \\ &= \Big((v^{(i)}(\tau))^\top R \otimes \Psi(x(\tau))^\top\Big)\mathrm{vec}(\hat{U}_i^\top). \end{aligned} \tag{63}$$

Substituting this into (62) gives the linear regression form $\tilde{\Theta}_i(t)\,[\mathrm{vec}(\hat{C}_i); \mathrm{vec}(\hat{U}_i^\top)] = \tilde{E}_i(t)$ with $\tilde{\Theta}_i(t), \tilde{E}_i(t)$ defined in (53)–(54). Stacking N samples yields (51). □

The least-square problem can be solved as detailed in [12, 16, 32]. It is noted that (51) implies a condition $N > N_1 + mN_2$ on the number of data samples that must be collected for a feasible solution to exist [1, 12]. To guarantee input saturation during the exploration and learning phases of the RL implementation, the actor network needs to be defined as follows:

$$u(x) = \rho\, \Gamma\big(U\, \Psi(x)\big) \tag{64}$$

where $\Gamma(\cdot)$ is a continuous one-to-one, bounded, integrable function of class $\mathcal{C}_p$ ($p \geq 1$) with $\Gamma(0) = 0$. Furthermore, ρ is a symmetric input saturation bound such that $|u| \leq \rho$. For systems with multiple inputs, the function $\Gamma(\cdot)$ is applied element-wise to the control inputs. For multiple inputs, saturation can be handled component-wise by replacing the scalar bound ρ with a vector $\rho \in \mathbb{R}^m_{>0}$ and enforcing $-\rho \leq u \leq \rho$ element-wise. The end-to-end tunable safe RL framework is presented in Algorithm 2 below.

Remark 10. While the online safety filter (32) requires a control-affine model (or an identified approximation) to

evaluate Lie derivatives, the off-policy policy-iteration update below can be implemented in a data-driven manner and does not require explicit model knowledge.

A. Computational Complexity and Practical Implementability

CLF/CBF-QP in (32) optimizes over $m+1$ decision variables ($u_{QP} \in \mathbb{R}^m$ and slack $\delta \in \mathbb{R}$) subject to $2m+2$ linear constraints ($2m$ saturation bounds plus one CLF and one CBF constraint). Thus, the per-step QP complexity scales primarily with the input dimension m (rather than the state dimension n); the low decision dimension (m+1) enables reliable real-time execution in typical embedded control settings as established in other studies as well [2]. On the other hand, off-policy approach involves actor-critic update that solves a linear least-squares regression with $p = N_1 + mN_2$ unknown parameters (critic size N_1, actor size N_2). Using QR factorization or normal equations, the computational cost per policy iteration is $O(Np^2 + p^3)$. For the experiments in Sec. VII, $N_1 = 22$, $N_2 = 9$, and $m = 1$, yielding $p = 31$. Regarding convergence, while a general analytical upper bound on the number of policy iteration steps is not available for general nonlinear systems, we state the iteration counts under the used stopping criterion.

Remark 11 (Guideline - parameter selection)**.** We typically fix ξ from the probing signal bound and choose ϵ_0 as the smallest value such that (32) remains feasible for states near ∂C under $\|e_u\|_\infty \leq \xi$. Then, we choose large (but not excessively large) value for $\epsilon_{\max}$ to limit the maximum enlargement $\gamma_T(h, \xi)$ induced. Finally, we choose τ by prescribing a transition margin $h_{\mathrm{tr}} > 0$ and a fraction $\eta \in (0, 1)$ such that $\epsilon(h_{\mathrm{tr}}) = \epsilon_0 + \eta(\epsilon_{\max} - \epsilon_0)$, which yields $\tau = -\frac{\sigma(h_{\mathrm{tr}})}{\ln(1-\eta)}$.

VII. Simulations

A. Inverted Pendulum

Consider an inverted pendulum with dynamics $\forall t \geq 0$ given by,

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ \frac{3g}{2l} \sin x_1 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{3}{ml^2} u \end{bmatrix} \tag{65}$$

where x_1 and x_2 are the angle and angular velocities of the pendulum respectively, and u is the control input (torque) with $m = 1$kg, $g = 9.81$m/s^2, $l = 1$m. The safety bounds for x_2 are taken as $\mathcal{C} = \{x \in \mathbb{R}^n \mid -5 \leq x_2 \leq 3\}$ and the utility (reward) function (33) is chosen with $Q = \texttt{diag([0.1,0.01])}$, $R = 0.001$ and $|u| \leq 5$. The critic and actor basis functions are selected as,

$$\begin{aligned} \phi(x) = \Big[& x_1^2, x_2^2, x_1x_2, x_1^4x_2^4, x_1^3x_2^2, x_1^2x_2^2, x_1x_2^3, x_1^6, x_2^6, \\ & x_1^5x_2, x_1^4x_2^2, x_1^3x_2^3, x_1^2x_2^4, x_1x_2^5, x_1^8x_2^2, x_1^7x_2, x_1^6x_2^2, x_1^4x_2^4, \\ & x_1^3x_2^5, x_1^2x_2^6, x_1x_2^7, x_1^{10}x_2^{10} \Big]^\top, \end{aligned} \tag{66}$$

$$\psi(x) = \Big[x_1, x_2, x_1x_2, x_1^2x_2^2, x_1^3, x_2^3, x_1^2x_2^3, x_1^4, x_2^4 \Big]^\top. \tag{67}$$

Algorithm 2 Off-policy Safe RL with Input Saturation

Require: Baseline feedback $u_0 : \Omega \to \mathbb{R}^m$; probing signal $e_u(t)$; QP safety filter (32); basis functions Φ, Ψ.

1: **Initialization:** Define the initial admissible safe policy for PI as $u^{(0)}(x) \leftarrow u_0(x) + u^\star_{QP}(x, u_0(x))$.
2: **procedure** Data Collection
3: Apply the pre-filter behavior input $u_{\mathrm{noisy}}(t) \leftarrow u_0(x(t)) + e_u(t)$.
4: Solve (32) point-wise to obtain $u_{QP}(t) \leftarrow u^\star_{QP}(x(t), u_{\mathrm{noisy}}(t))$.
5: Apply $u_{\mathrm{applied}}(t) \leftarrow u_{\mathrm{noisy}}(t) + u_{QP}(t)$ to the system.
6: Collect regression data $\tilde{\Theta} = [\theta_1, \ldots, \theta_N]$, $\tilde{E} = [e_1, \ldots, e_N]$ until $N \geq N_1 + mN_2$.
7: **procedure** Off-policy Safe Policy Iteration (Algorithm 1)
8: $i \leftarrow 0$.
9: **repeat**
10: Form $v^{(i)}(t_k) \leftarrow u_{\mathrm{applied}}(t_k) - u^{(i)}(x(t_k))$ from data.
11: Build $\tilde{\Theta}_i^N, \tilde{E}_i^N$ using (53)–(54).
12: $\hat{\theta}_i \leftarrow \left(\tilde{\Theta}_i^N \tilde{\Theta}_i^{N\top}\right)^{-1} \tilde{\Theta}_i^N \tilde{E}_i^{N\top}$.
13: Extract $\hat{C}_i$ and $\hat{U}_i$ from $\hat{\theta}_i$, set $u^{(i+1)}(x) \leftarrow \rho \tanh(\hat{U}_i \Psi(x))$.
14: $i \leftarrow i + 1$.
15: **until** $\|\widehat{W}_i - \widehat{W}_{i-1}\| \leq 10^{-12}$

The initial stabilizing policy is computed using by solving the QP-CLF optimization. The probing noise (excitation signal) added to the control input is given $\forall t \geq 0$ by:

$$e(t) = \sum_{k=1}^{K} a_k \, \omega_k \, \sin(\zeta_k t), \tag{68}$$

where $a = 8.5$, $\zeta = [1, 3, 7, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29]$, $K = \mathrm{length}(\zeta)$ and ω is a random variable with a zero mean Gaussian distribution $\omega \sim \mathcal{N}(0, \sigma_1)$ with $\sigma_1^2 = 1$, truncated between $[-3\sigma_1, 3\sigma_1]$.

1) Safe Exploration: Algorithm 2 is implemented in a similar fashion as [12] (due to space restrictions, classical details are omitted here).

$\epsilon_{\max}$	τ
50	0.5,2.5,5.0
100	0.5,2.5,5.0
500	0.5,2.5,5.0
750	0.1,0.5,2.5,5.0
1000	0.1,0.5,2.5,5.0

TABLE I: Parameter settings.

Figure 1 compares the exploration-phase evolution of x_2 under the classical CBF filter [12] and under the proposed ISSf framework with different $\epsilon(h(x))$ schedules (constant, reciprocal $1/h(x)$, and the proposed ϵ-tuning law). The classical CBF preserves safety but exhibits uniformly conservative behavior leading to no preferable exploration near safety boundaries but

keeps x_2 well inside the admissible set. Fixed or state based (reciprocal) $\epsilon(h(x))$ can reduce this conservatism, but may yield inconsistent behavior across the safe set (overly timid in the interior and/or insufficiently protective near the boundary). However, the proposed ϵ-tuning law adaptively modulates ϵ to promote boundary-proximal excitation while maintaining constraint satisfaction leading to preferable exploration near boundaries yielding richer data collection whilst maintaining safety under the chosen high magnitude excitation.

Further simulations are conducted with a set of varying $\epsilon_{\max}$ and τ values, whilst using $\mathcal{K}_\infty^e$ choice $\alpha_4(s) = k_4 s$ with $k_4 = 10$. Figure 2 illustrates the effect of varying $\epsilon_{\max}$ with a fixed $\tau = 2.5$. Increasing $\epsilon_{\max}$ relaxes the ISSf-Exp-CBF constraint away from the boundary (since $\|L_g h\|^2/\epsilon(h)$ decreases), thereby reducing conservatism and enabling more boundary-proximal exploration; excessively large $\epsilon_{\max}$ may, however, reduce robustness margins under aggressive excitation. Conversely, smaller $\epsilon_{\max}$ yields a more conservative evolution with a larger safety margin, at the cost of reduced excitation near the constraint. Figure 3 presents the results with a varying τ and a fixed $\epsilon_{\max} = 50$. The parameter τ governs how quickly $\epsilon(h(x))$ adapts: smaller τ produces faster gain variation (sharper transients near constraint activation), whereas larger τ slows the adaptation and results in smoother, more conservative transients. Fig. 3 shows that τ determines the rate at which ϵ increases or decreases to reach a specific $\epsilon_{\max}$, thereby influencing the level of conservatism.

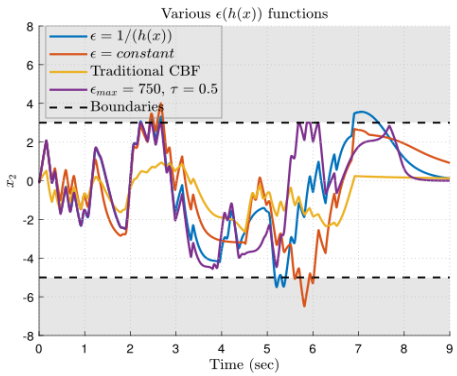

Fig. 1: Exploration-phase evolution of x_2 for different $\epsilon(h(x))$ schedules and the classical CBF baseline.

Figure 4 shows the evolution of the input profile where the safety filter enforces the saturation constraint while applying only the minimal correction needed to satisfy the CLF/CBF inequalities.

2) Safe learning: For safe learning, B_λ (cf. (34)) is chosen with damping coefficient $\lambda = 2.5$. Post exploration (safe data collection), the learning of actor-critic neural networks training takes around 17 iteration steps as shown in Fig. 6, converging to optimality within small number of steps. Fig. 7 shows value function surface at two different iterations illustrating the monotonic nature of value function as claimed in Lemma 5. Figure 5 compares the closed-loop state trajectories under the same high-magnitude, persistently exciting exploration noise. Without any safety filter during data collection (baseline: noisy behavior policy with barrier-augmented learning objective

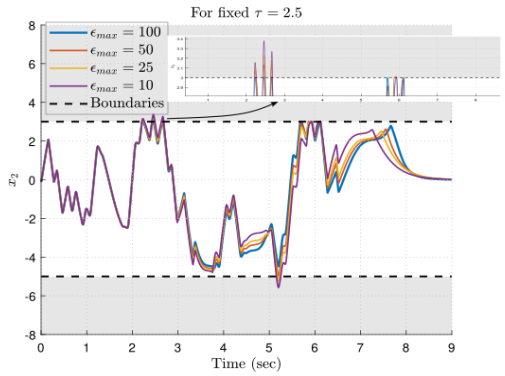

Fig. 2: The evolution of state x_2 under the effect of varying $\epsilon_{\max}$.

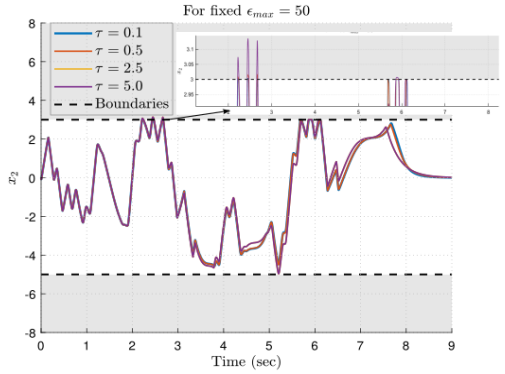

Fig. 3: The evolution of state x_2 under the effect of varying τ.

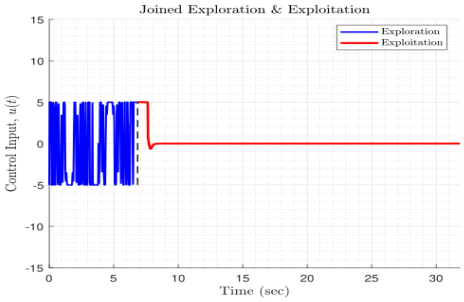

Fig. 4: The evolution of the control input under saturation.

only), the exploration input drives the system outside the prescribed safety bounds, yielding constraint violations during exploration. The classical CLF-CBF-QP safety filter (without ISSf/TISSf robustness terms) prevents violations but remains conservative, keeping trajectories well inside the safe set and thereby limiting boundary-proximal excitation. In contrast, the proposed ISSf-based filter preserves constraint satisfaction while enabling richer, boundary-aware exploration under the same noise level.

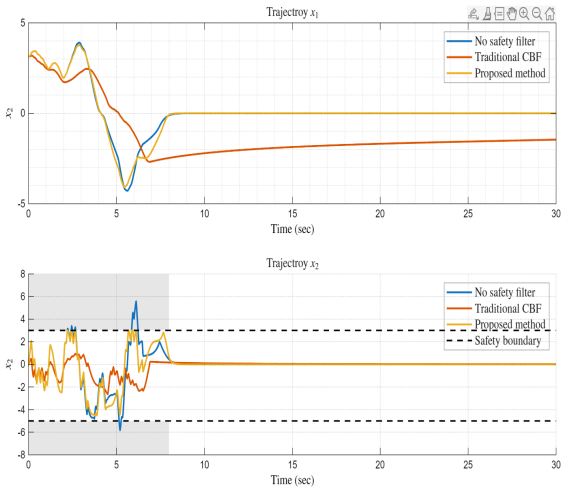

Fig. 5: Example 1: System trajectories under identical high-magnitude exploration noise: no safety filter, traditional CBF (without ISSf), and the proposed ISSf-based filter

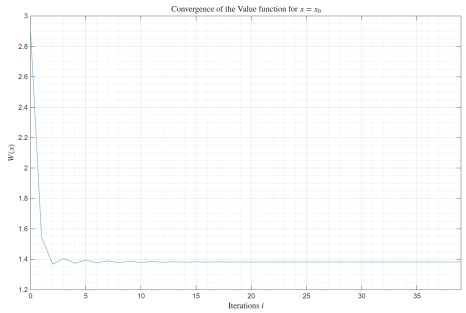

Fig. 6: Value function vs. number of iterations

B. Jet Engine

Consider the following jet engine surge and stall dynamics [11, 12] $\forall t \geq 0$:

$$\begin{aligned} \dot{x}_1 &= -0.35x_1^2 - 0.35x_1(2x_1 + x_2^2) \\ \dot{x}_2 &= -1.4x_2^2 - 0.5x_2^3 - 3x_1x_2 - 3x_1 - u \end{aligned} \tag{69}$$

where x_1 is the normalized rotating stall amplitude, x_2 is the deviation of the annulus-averaged flow, and the control input u is the deviation of the plenum pressure rise. The safe set for x_2 is given by $\mathcal{C} = \{x_2 | -1.1 < x_2 < 0.45\}$. The user-defined matrices are selected as $Q = \texttt{diag([20,10])}$, $R = 0.3$. The control input is saturated within the bounds $|u| \leq 10$. The class $\mathcal{K}_\infty$ function is taken as $\alpha_4 = 20$. The basis functions are selected as,

$$\begin{aligned} \phi(x) = \Big[&x_1^2, x_2^2, x_1x_2, x_1^4x_2^4, x_1^3x_2^2, x_1^2x_2^2, x_1x_2^3, x_1^6, x_2^6, \\ &x_1^5x_2, x_1^4x_2^2, x_1^3x_2^3, x_1^2x_2^4, x_1x_2^5, x_1^8x_2^2, x_1^7x_2, x_1^6x_2^2, x_1^4x_2^4, \\ &x_1^3x_2^5, x_1^2x_2^6, x_1x_2^7, x_1^{10}x_2^{10}\Big]^\top, \end{aligned} \tag{70}$$

$$\psi(x) = \Big[x_1, x_2, x_1x_2, x_1^2x_2^2, x_1^3, x_2^3, x_1^2x_2^3, x_1^4, x_2^4\Big]^\top. \tag{71}$$

1) Safe Exploration: The exploration noise is given by (68). The parameter values considered are shown in Table II. Figure 8 compares the exploration-phase evolution of x_2 under

$\epsilon_{\max}$	τ
50	0.5,2.5,5.0
100	0.5,2.5,5.0
500	0.5,2.5,5.0
750	0.1,0.5,2.5,5.0
1000	0.1,0.5,2.5,5.0

TABLE II: Parameter settings.

the classical CBF filter [12] and under ISSf-based filtering with three $\epsilon(h(x))$ schedules: (i) constant $\epsilon = 120$, (ii) reciprocal $\epsilon = 1/h(x)$, and (iii) the proposed ϵ-tuning law (e.g., $\epsilon_{\max} = 750,\ \tau = 1$). The classical CBF assures safety but yields conservative (non rich) exploration. In contrast, fixed or reciprocal for $\epsilon(h(x))$ choices can produce larger excursions yet exhibit boundary violations under aggressive exploration noise, indicating insufficient robustness near danger zones. The proposed ϵ-tuning adapts ϵ to strengthen correction as $h(x) \to 0$ while relaxing in the interior, thereby enabling boundary-aware (richer) exploration without safety violations.

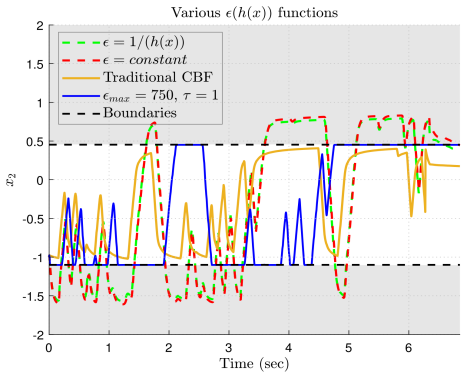

Fig. 8: Exploration-phase evolution of x_2 under the classical CBF baseline and different $\epsilon(h(x))$ schedules.

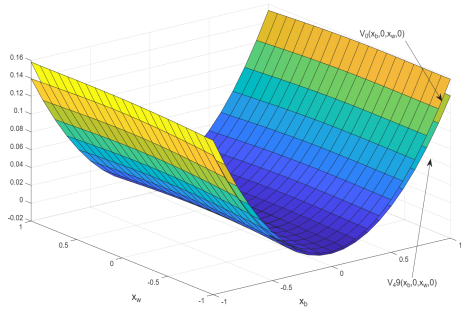

Fig. 7: Value function surface at two distinct iterative steps during learning

Figure 9 illustrates the influence of $\epsilon_{\max}$ on x_2 for fixed $\tau = 0.5$ (with $\alpha_4 = 10$). Increasing $\epsilon_{\max}$ diminishes the admissible range of the adaptive gain when constraints become active, resulting in weaker QP correction and a less conservative response leading to boundary proximal exploration (i.e., trajectories remain preferably near the safety bounds). Notably,

the trajectories largely overlap away from constraint activation, indicating that $\epsilon_{\max}$ primarily shapes the behavior near the safety boundaries. Figure 10 shows the effect of the adaptation time-scale τ for fixed $\epsilon_{\max} = 1000$ (with $\alpha_4 = 10$). Smaller τ yields faster ϵ adaptation, tightening the filter more rapidly near the boundary and relaxing it more quickly in the interior; this produces sharper transients and increased switching during exploration. Larger τ slows the gain variation, leading to smoother but more conservative transient behavior. Across the tested values, τ mainly tunes transient aggressiveness rather than the overall safe-set invariance.

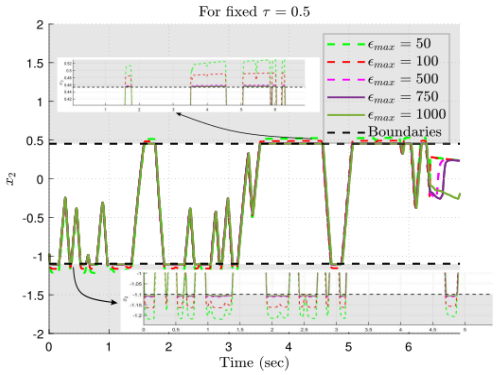

Fig. 9: The evolution of state x_2 under the effect of $\epsilon_{\max}$.

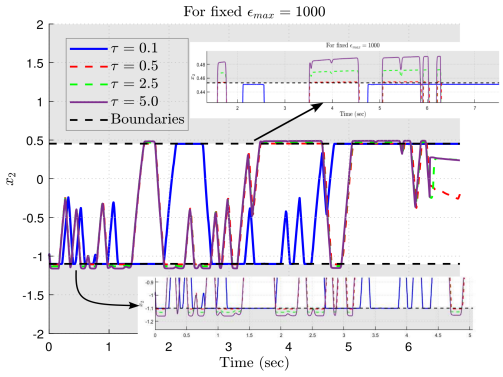

Fig. 10: The evolution of state x_2 under the effect of varying τ.

2) Safe learning: Actor-critic NNs take around 7-8 iterative steps to get trained. Figure 11 compares the closed-loop trajectories under identical high-magnitude exploration noise for three cases: (i) no safety filter, (ii) a classical CBF filter (no ISSf condition), and (iii) the proposed ISSf-based filter. Without safety filtering, the exploration input drives the system outside the prescribed safety bounds (constraint violations during exploration). The classical CBF prevents violations but is conservative, keeping the trajectories away from the boundaries and limiting boundary-proximal excitation. In contrast, the proposed approach preserves constraint satisfaction while allowing richer, boundary-aware exploration under the same noise level, with regulation performance remaining comparable.

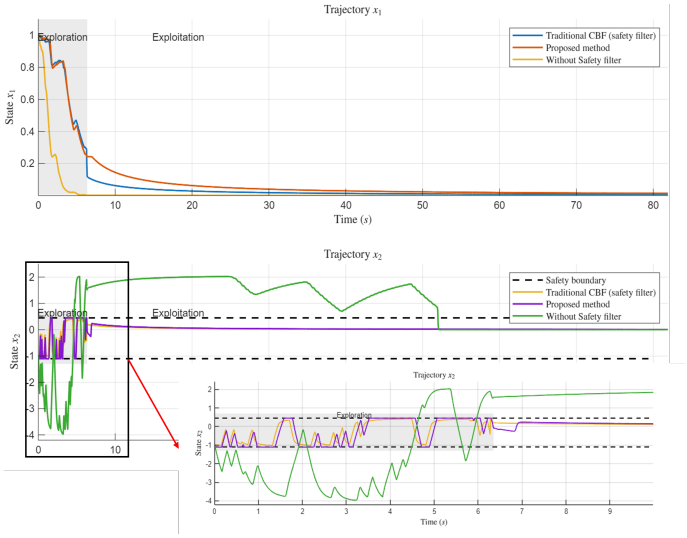

Fig. 11: Example 2: System trajectories under identical high-magnitude exploration noise: no filter, classical CBF, and proposed ISSf-based filter.

VIII. Conclusion

The proposed novel approach guarantees safe initialization, safe exploration as well as safe learning under input saturation limits encouraging safety boundary-proximal exploration leading to richer data collection. The proposed approach addresses conservatism issue during exploration through a novel ϵ-tuning law that dynamically adjusts conservatism, strictly enforcing safety near boundaries while allowing richer exploration deeper within the safe region, enabling precise control over exploration conservatism. Simulation results demonstrate that this procedure guarantees system safety under aggressive exploration without sacrificing exploratory richness. Input saturation are suitably modeled and considered within the cost function along with reciprocal barrier functions to guarantee safety during learning. Neural networks are used for approximating the value function and control policy. The novel analytically derived closed-form least-square expression under input saturation enables the application of the proposed approach in an off-policy manner. Future work includes considering unmatched disturbances in a rigorous manner. Further, proposed framework is compatible with data-driven system identification and model learning approaches. For example, bounded modeling/identification error can be treated as an additional disturbance term and absorbed into the disturbance/noise bound. The latter will be explored to alleviate model dependence.

References

[1] Murad Abu-Khalaf and Frank L Lewis. "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach". In: *Automatica* 41.5 (2005), pp. 779–791.

[2] Anil Alan, Andrew J Taylor, Chaozhe R He, Aaron D Ames, and Gábor Orosz. "Control barrier functions and input-to-state safety with application to automated vehicles". In: *IEEE Transactions on Control Systems Technology* 31.6 (2023), pp. 2744–2759.

[3] Anil Alan, Andrew J Taylor, Chaozhe R He, Gábor Orosz, and Aaron D Ames. "Safe controller synthesis with tunable input-to-state safe control barrier functions". In: *IEEE Control Systems Letters* 6 (2021), pp. 908–913.

[4] Aaron D Ames, Xiangru Xu, Jessy W Grizzle, and Paulo Tabuada. "Control barrier function based quadratic programs for safety critical systems". In: *IEEE Transactions on Automatic Control* 62.8 (2016), pp. 3861–3876.

[5] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. "Safe learning in robotics: From learning-based control to safe reinforcement learning". In: *Annual Review of Control, Robotics, and Autonomous Systems* 5 (2022), pp. 411–444.

[6] Christian Feller and Christian Ebenbauer. "Continuous-time linear MPC algorithms based on relaxed logarithmic barrier functions". In: *IFAC Proceedings Volumes* 47.3 (2014), pp. 2481–2488.

[7] Randy A Freeman and Petar V Kokotovic. "Inverse optimality in robust stabilization". In: *SIAM journal on control and optimization* 34.4 (1996), pp. 1365–1391.

[8] Mrdjan Jankovic. "Combining control Lyapunov and barrier functions for constrained stabilization of nonlinear systems". In: *2017 American control conference (ACC)*. IEEE. 2017, pp. 1916–1922.

[9] Mrdjan Jankovic. "Robust control barrier functions for constrained stabilization of nonlinear systems". In: *Automatica* 96 (2018), pp. 359–367.

[10] Mayank Shekhar Jha and Bahare Kiumarsi. "Off-policy safe reinforcement learning for nonlinear discrete-time systems". In: *Neurocomputing* 611 (2025), p. 128677.

[11] Yu Jiang and Zhong-Ping Jiang. "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems". In: *IEEE Transactions on Neural Networks and Learning Systems* 25.5 (2014), pp. 882–893.

[12] Soha Kanso, Mayank Shekhar Jha, and Didier Theilliol. "Off-policy model-based end-to-end safe reinforcement learning". In: *International Journal of Robust and Nonlinear Control* 34.4 (2024), pp. 2806–2831.

[13] HK Khalil. *Nonlinear systems*. 2002.

[14] Bahare Kiumarsi and Frank L Lewis. "Actor–critic-based optimal tracking for partially unknown nonlinear discrete-time systems". In: *IEEE transactions on neural networks and learning systems* 26.1 (2014), pp. 140–151.

[15] Shishir Kolathaya and Aaron D Ames. "Input-to-state safety with control barrier functions". In: *IEEE control systems letters* 3.1 (2018), pp. 108–113.

[16] Frank L Lewis, Draguna Vrabie, and Kyriakos G Vamvoudakis. "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers". In: *IEEE Control Systems Magazine* 32.6 (2012), pp. 76–105.

[17] S Edward Lyshevski. "Optimal control of nonlinear continuous-time systems: design of bounded controllers via generalized nonquadratic functionals". In: *Proceedings of the 1998 American Control Conference. ACC (IEEE Cat. No. 98CH36207)*. Vol. 1. IEEE. 1998, pp. 205–209.

[18] Zahra Marvi and Bahare Kiumarsi. "Barrier-Certified Learning-Enabled Safe Control Design for Systems Operating in Uncertain Environments". In: *IEEE/CAA Journal of Automatica Sinica* 9.3 (2021), pp. 437–449.

[19] Zahra Marvi and Bahare Kiumarsi. "Safe reinforcement learning: A control barrier function optimization approach". In: *International Journal of Robust and Nonlinear Control* 31.6 (2021), pp. 1923–1940.

[20] Zahra Marvi and Bahare Kiumarsi. "Safety planning using control barrier function: A model predictive control scheme". In: *2019 IEEE 2nd Connected and Automated Vehicles Symposium (CAVS)*. IEEE. 2019, pp. 1–5.

[21] Marcel Menner and Eugene Lavretsky. "Translation of Nagumo's Foundational Work on Barrier Functions: On the Location of Integral Curves of Ordinary Differential Equations". In: *arXiv preprint arXiv:2406.18614* (2024).

[22] Amir Modares, Nasser Sadati, Babak Esmaeili, Farnaz Adib Yaghmaie, and Hamidreza Modares. "Safe reinforcement learning via a model-free safety certifier". In: *IEEE Transactions on Neural Networks and Learning Systems* 35.3 (2023), pp. 3302–3311.

[23] Hamidreza Modares, Frank L Lewis, and Mohammad-Bagher Naghibi-Sistani. "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks". In: *IEEE Transactions on neural networks and learning systems* 24.10 (2013), pp. 1513–1525.

[24] Mitio Nagumo. "Über die lage der integralkurven gewöhnlicher differentialgleichungen". In: *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series* 24 (1942), pp. 551–559.

[25] Chunbin Qin, Suyang Hou, Mingyu Pang, Zhongwei Wang, and Dehua Zhang. "Reinforcement learning-based secure tracking control for nonlinear interconnected systems: An event-triggered solution approach". In: *Engineering Applications of Artificial Intelligence* 161 (2025), p. 112243.

[26] Chunbin Qin, Xiaopeng Qiao, Jinguang Wang, Dehua Zhang, Yandong Hou, and Shaolin Hu. "Barrier-critic adaptive robust control of nonzero-sum differential games for uncertain nonlinear systems with state constraints". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 54.1 (2023), pp. 50–63.

[27] Chunbin Qin, Xiaopeng Qiao, Jinguang Wang, Dehua Zhang, and Shuyu Hu. "Observer Based Fault Tolerant Control Design for Saturated Nonlinear Systems with Full State Constraints via a Novel Event-Triggered Mechanism". In: *Engineering Applications of Artificial Intelligence* 161 (2025), p. 112221. DOI: 10.1016/j.engappai.2025.112221.

[28] Muhammad Zakiyullah Romdlony and Bayu Jayawardhana. "On the new notion of input-to-state safety". In: *2016 IEEE 55th conference on decision and control (CDC)*. IEEE. 2016, pp. 6403–6409.

[29] George N Saridis and Chun-Sing G Lee. "An approximation theory of optimal control for trainable manipulators". In: *IEEE Transactions on systems, Man, and Cybernetics* 9.3 (1979), pp. 152–159. ISSN: 0018-9472.

[30] Eduardo D Sontag. "A 'universal'construction of Artstein's theorem on nonlinear stabilization". In: *Systems & control letters* 13.2 (1989), pp. 117–123.

[31] Aviv Tamar, Huan Xu, and Shie Mannor. "Scaling up robust MDPs by reinforcement learning". In: *arXiv preprint arXiv:1306.6189* (2013).

[32] Asma Al-Tamimi, Frank L Lewis, and Murad Abu-Khalaf. "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 38.4 (2008), pp. 943–949.

[33] Kyriakos G Vamvoudakis and Frank L Lewis. "Multi-player non-zero-sum games: Online adaptive learning solution of coupled Hamilton–Jacobi equations". In: *Automatica* 47.8 (2011), pp. 1556–1569.

[34] Hao Wang, Jiahu Qin, and Zhen Kan. "Shielded planning guided data-efficient and safe reinforcement learning". In: *IEEE Transactions on Neural Networks and Learning Systems* 36.2 (2024), pp. 3808–3819.

[35] Li Wang, Aaron D Ames, and Magnus Egerstedt. "Safety barrier certificates for collisions-free multirobot systems". In: *IEEE Transactions on Robotics* 33.3 (2017), pp. 661–674.

[36] Peter Wieland and Frank Allgöwer. "Constructive safety using control barrier functions". In: *IFAC Proceedings Volumes* 40.12 (2007), pp. 462–467.

[37] Yujie Yang, Yuxuan Jiang, Yichen Liu, Jianyu Chen, and Shengbo Eben Li. "Model-free safe reinforcement learning through neural barrier certificate". In: *IEEE Robotics and Automation Letters* 8.3 (2023), pp. 1295–1302.

[38] Dehua Zhang, Yuchen Wang, Lei Meng, Jiayuan Yan, and Chunbin Qin. "Adaptive critic design for safety-optimal FTC of unknown nonlinear systems with asymmetric constrained-input". In: *ISA transactions* 155 (2024), pp. 309–318.

[39] Lingzhi Zhang, Runze Lin, Lei Xie, Wei Dai, and Hongye Su. "Event-triggered constrained optimal control for organic Rankine cycle systems via safe reinforcement learning". In: *IEEE Transactions on Neural Networks and Learning Systems* 35.5 (2022), pp. 7126–7137.

[40] Yuxiang Zhang, Xiaoling Liang, Dongyu Li, Shuzhi Sam Ge, Bingzhao Gao, Hong Chen, and Tong Heng Lee. "Barrier Lyapunov function-based safe reinforcement learning for autonomous vehicles with optimized backstepping". In: *IEEE Transactions on Neural Networks and Learning Systems* 35.2 (2022), pp. 2066–2080.

[41] Zhehua Zhou, Ozgur S Oguz, Marion Leibold, and Martin Buss. "Learning a low-dimensional representation of a safe region for safe reinforcement learning on dynamical systems". In: *IEEE Transactions on Neural Networks and Learning Systems* 34.5 (2021), pp. 2513–2527.