Lipschitz-Aware Exploration for Safety in Reinforcement Learning

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

Abstract— **This paper develops a novel Lipschitz-aware safe exploration framework for reinforcement learning in environments with abrupt, unmodeled safety variations. Local Lipschitz constants of the safety function are estimated online using kernel density estimation (KDE), providing a data-driven measure of rapid changes in the safety landscape. These estimates are incorporated into a robust quadratic program (QP) with Lipschitz-aware control barrier function (CBF) constraints, yielding a safe exploration law that guarantees forward invariance of an enlarged safe set under probing noise. The exploration mechanism is then coupled with a safety-aware learning stage to obtain a unified safe RL framework. Simulations on an inverted pendulum illustrate the efficacy of the proposed approach.**

I. Introduction

Safety in control systems is increasingly addressed using Control Barrier Functions (CBFs), which ensure forward invariance of safe sets under control actions [1, 2]. To account for bounded disturbances, Input-to-State Safety (ISSf) was introduced in [3], leading to ISSf-CBFs. While effective, these can be overly conservative. Recent works [4, 5] propose Tunable ISSf-CBFs (TISSf-CBFs), which allow tuning the invariant set to closely approximate the undisturbed safe set, reducing conservatism without significantly impacting the performance.

Reinforcement Learning (RL) offers a model-free paradigm for learning stabilizing and optimal control policies [6], including online solutions to the Hamilton–Jacobi–Bellman (HJB) equation via Policy Iteration (PI) [7–9]. Safe RL combines RL with safety guarantees, often using CBFs to enforce safety during both exploration and learning [10–13]. Off-policy PI algorithms are especially attractive as they separate exploration (via a behavioral policy) and exploitation (via a target policy) [14]. However, most works assume an initial admissible controller and do not address safety during exploration.

Recent work in [12] introduced an off-policy PI framework that ensures safety across initialization, exploration, and learning by using Quadratic Program (QP) controllers under joint CBF and Control Lyapunov Function (CLF) constraints. However, such QP-based controllers can become infeasible in the presence of unmodeled or rapidly varying nonlinearities. To address feasibility loss, [15] proposed a feasible set reshaping approach. Separately, [16] used Kernel Density Estimation (KDE) to estimate the Lipschitz constant of unknown nonlinearities, though this was not applied to safe exploration.

Most safe RL methods assume smooth variation in safety limits, but real-world environments often undergo abrupt changes (e.g., obstacles, lane shifts), causing sharp transitions in the safety function. Exploring near safety boundaries is crucial for data efficiency, yet aggressive exploration noise in such settings can induce rapid, unmodeled variations in the safety function. This may render QP controllers infeasible. Thus, balancing exploration aggressiveness with conservatism becomes critical to ensure safety during exploration in uncertain, dynamic environments.

Recently, the Tunable ISSf strategy was applied to safe exploration in [17, 18], where exploration is made highly conservative near safety boundaries and remains non-conservative within the interior of the safe set. However, these works do not account for abrupt environmental changes or variations in the Lipschitz constant of the safety function.

Contributions: This paper: i) formulates safe exploration as a Lipschitz-aware robust QP with CBF constraints, using KDE to estimate local Lipschitz constants online; ii) introduces an enlarged safe set that incorporates this local Lipschitz information and guarantees robust forward invariance during exploration; and iii) interfaces the resulting exploration module with a safety-aware learning stage to obtain an integrated RL framework for rapidly varying environments.

Notation: Standard notions of class $\mathcal{K}$, $\mathcal{K}_\infty$, and extended class $\mathcal{K}_\infty^e$ functions are used. For any essentially bounded signal $d : [0,\infty) \to \mathbb{R}^p$, let supremum be denoted as $\|d\|_\infty := \sup_{t\geq 0} \|d(t)\|$. We use $\mathcal{U} \subseteq \mathbb{R}^m$ for instantaneous control values and $\mathfrak{U}$ for measurable input signals $u : [0,\infty) \to \mathbb{R}^m$. For any differentiable scalar function φ, $L_f\varphi(x) := \nabla\varphi(x)^\top f(x)$ and $L_g\varphi(x) := \nabla\varphi(x)^\top g(x)$.

Structure: Section II provides an overview of the background concepts. Section III outlines the problem formulation, while Section IV introduces the novel propositions on Lipschitz-aware safe exploration. The safe learning approach is presented in Section V, followed by a simulation study in Section VI. Lastly, Section VII concludes the paper and discusses directions for future research.

II. Background

Consider an affine-in-control nonlinear system for all $t \geq 0$:

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

[1] M. S. Jha, S. Marthi, and D. Theilliol are with Universite de Lorraine, CNRS, CRAN, F-54000 Nancy, France. *MSJ is Corresponding Author. 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 by: French National Research Project ANR SOS (ANR-23-IAS2-0003) as well as NSF under grant Nos. CAREER CPS-1851588, CPS-2227185, S&AS-1849198, and SLES-2415479.

where $x \in \Omega \subseteq \mathbb{R}^n$ is the state, $u \in \mathcal{U} \subseteq \mathbb{R}^m$ is the control input. We assume $f \in C^1$, g is locally Lipschitz on Ω, $f(0) = 0$, and that the system is stabilizable on the compact set Ω. Hence, $x = 0$ is an equilibrium under $u = 0$.

A. Optimal Control

Here, the objective is to find a policy $u(\cdot)$ minimizing the infinite-horizon cost

$$V(x) = \int_0^\infty r(x,u)\,\mathrm{d}t, \qquad r(x,u) = q(x) + u^\top R u, \quad (2)$$

where $q : \Omega \to \mathbb{R}_{\geq 0}$ is positive definite and $R \in \mathbb{R}^{m\times m}$ is positive definite. A policy is admissible if it stabilizes the system and yields finite cost. Assuming $V \in C^1$, the value function satisfies:

$$\nabla V^\top (f(x) + g(x)u) = -r(x,u). \quad (3)$$

Substituting the optimal policy $u^\star = -\frac{1}{2}R^{-1}g^\top(x)\nabla V^\star$, this yields the Hamilton–Jacobi–Bellman (HJB) equation:

$$\begin{aligned} H(V^\star)(x) &:= \nabla V^{\star\top} f(x) + q(x) \\ -\tfrac{1}{4}\nabla V^{\star\top} &g(x)R^{-1}g^\top(x)\nabla V^\star = 0. \end{aligned} \quad (4)$$

In general, closed-form solutions to (4) are intractable, motivating iterative solutions such as Policy Iteration (PI) involving evaluation and improvement steps [6]. Classical formulations do not incorporate safety, which is addressed in Safe RL frameworks [11–13].

B. Safety

Let the safe set be

$$\mathcal{C} := \{x \in \mathbb{R}^n : h(x) \geq 0\}, \quad (5)$$

where $h : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Safety is understood as forward invariance of $\mathcal{C}$. In the zeroing-barrier formulation, the safe-set defining function h itself serves as the barrier function. Specifically, h is a ZCBF if there exists $\alpha_4 \in \mathcal{K}^e_\infty$ such that

$$\sup_{u\in\mathcal{U}} \left[L_f h(x) + L_g h(x) u\right] \geq -\alpha_4(h(x)), \qquad \forall x \in \mathcal{D}. \quad (6)$$

We also use the standard CLF and ISSf notions from [1, 3–5] for admissible initialization and robust safety interpretation.

III. Problem Formulation

Consider the system (1) during exploration:

$$\dot{x} = f(x) + g(x)\big(u + e_u(t)\big), \qquad t \geq 0, \quad (7)$$

where $e_u : [0,\infty) \to \mathbb{R}^m$ is a bounded, piecewise-continuous probing signal satisfying $\|e_u\|_\infty \leq \xi$, with $\|e_u\|_\infty := \sup_{t\geq 0}\|e_u(t)\|$. Although injected deliberately for data collection, $e_u(t)$ is treated in the analysis as a matched disturbance.

Even when $f(x)$, $g(x)$, and a nominal safety function $h(x) \in C^1$ are known, abrupt environmental changes can induce localized steep variations in $h(x)$, effectively shifting the safe set boundary. Such variations are naturally characterized by large local Lipschitz quotients $|h(x_1) - h(x_2)|/\|x_1 - x_2\|$. Since these effects may not be captured by a fixed nominal model of $h(x)$, we use a data-driven estimate of the local Lipschitz constant. In safe RL, exploration near the boundary is informative but probing noise combined with rapidly varying safety limits can distort $h(x)$ and lead to infeasibility of QP-based safety filters. A large estimate $\widehat{\eta}_h$ therefore serves as an indicator of abrupt variations and motivates a Lipschitz-aware safety mechanism.

Remark 1. Although $h(x) \in C^1$ is assumed for controller synthesis, the safety landscape encountered during exploration may still exhibit localized steep variations. The estimate $\widehat{\eta}_h$ is used as a real-time sensitivity indicator to capture such effects without requiring an explicit perturbation model. □

IV. Safe Exploration

Classical CBF designs assume that the boundary $\partial\mathcal{C}$ evolves smoothly. However, abrupt environmental transitions (e.g., dynamic obstacles) cause rapid, localized shifts in $h(x)$, which may not be captured analytically. The local Lipschitz constant estimated via KDE provides a data-driven measure of this variability, allowing the safe set to adapt spatially, rather than relying on global or fixed bounds. First, the KDE based approach developed in [16] is presented to learn Lipschitz constant associated with unknown nonlinearity manifesting in $h(x)$ under rapid environmental variations.

A. Local Lipschitz Estimation

Since $h \in C^1(\Omega)$ and Ω is compact, h is globally Lipschitz on Ω; i.e., there exists $\eta_h^\star > 0$ such that

$$|h(x_1) - h(x_2)| \leq \eta_h^\star \|x_1 - x_2\|, \qquad \forall x_1, x_2 \in \Omega.$$

To capture abrupt spatial variations during exploration, however, we estimate a local Lipschitz constant from a moving data buffer $D_h(t_k) = \{(x_i, h(x_i))\}_{i=k-N_h+1}^{k}$. For each pair (x_i, x_j) in the buffer, define the Lipschitz quotient $\ell^h_{ij} := \frac{|h(x_i)-h(x_j)|}{\|x_i - x_j\|}$. Let $\{\ell^h_r\}_{r=1}^{n_\ell}$ collect these quotients. Their empirical density is estimated via KDE as

$$\widehat{\rho}_h(\ell) = \frac{1}{n_\ell b_h}\sum_{r=1}^{n_\ell} \mathcal{K}\left(\frac{\ell - \ell^h_r}{b_h}\right), \quad (8)$$

where $\mathcal{K}$ is a Gaussian kernel and $b_h > 0$ is the bandwidth. Given a density threshold $\beta_h > 0$, define $S_h(t_k) := \{\ell \geq 0 : \widehat{\rho}_h(\ell) \geq \beta_h\}$. The scalar Lipschitz estimate used by the controller at time t_k is

$$\widehat{\eta}_h(t_k) := \max S_h(t_k). \quad (9)$$

This estimate is then held constant over the interval $[t_k, t_{k+1})$. The buffer $D_h(t_k)$ is updated at every sampling instant t_k with sampling period T_s, and the estimate $\widehat{\eta}_h(t_k)$ is held constant over $[t_k, t_{k+1})$.

Remark 2. Under standard regularity conditions, $\widehat{\rho}_h(\ell)$ converges uniformly to the density of the Lipschitz quotients; see [16]. Here, the controller uses $\widehat{\eta}_h(t_k)$ as a data-driven local sensitivity indicator, with larger values indicating more rapidly varying or less accurately modeled safety boundaries.

The parameter β_h defines the density threshold for the support set $S_h(t_k)$ and should not be interpreted as a confidence level unless explicitly calibrated. □

The next section presents the novel propositions of this paper inspired from ISSf concepts in [3] and Tunable ISSf in [4], and extends them in safe exploration context.

B. Lipschitz-Aware Input-to-State Safety (L-ISSf)

Classical ISSf-CBF constructions enlarge the safe set according to a disturbance bound. Here, robustness must account for both the probing signal $e_u(t)$ and abrupt local variations in the safety region captured by $\widehat{\eta}_h$. We therefore construct a Lipschitz-aware enlarged safe set $\mathcal{C}_{\xi,L} \supseteq \mathcal{C}$, whose size depends on both the exploration budget ξ and the estimated local sensitivity of $h(x)$. At each sampling instant, $\widehat{\eta}_h$ is computed from the data buffer and held constant until the next update. We adopt the linear choice $\alpha_4(r) = \kappa r, \kappa > 0$, with corresponding enlargement map $\alpha_L(r) = r/\kappa$. Let $\eta_0 > 0$ be a nominal Lipschitz level and define the clipped estimate $\widetilde{\eta}_h := \max\{\widehat{\eta}_h, \eta_0\}$. Moreover, choose a constant $h_s > 0$ such that $h(x) + h_s > 0$ for all $x \in \Omega$. This shift guarantees that the robustness gain introduced below remains strictly positive over the operating domain. We now define the state- and Lipschitz-dependent robustness gain as

$$\bar{\epsilon}(h(x), \widehat{\eta}_h) = \epsilon_0 + k \, \frac{h(x) + h_s}{1 + \lambda \ln(\widetilde{\eta}_h/\eta_0)}, \tag{10}$$

where $\epsilon_0, k, \lambda > 0$ are design constants. The associated robust margin is

$$\delta(h(x), \xi, \widehat{\eta}_h) = \bar{\epsilon}(h(x), \widehat{\eta}_h) \frac{\xi^2}{4}, \tag{11}$$

and the enlarged barrier function is defined by

$$h_{\xi,L}(x, \xi, \widehat{\eta}_h) = h(x) + \frac{1}{\kappa}\, \delta(h(x), \xi, \widehat{\eta}_h). \tag{12}$$

The quantity $\bar{\epsilon}(h(x), \widehat{\eta}_h)$ acts as a tunable robustness gain, while $\delta(h(x), \xi, \widehat{\eta}_h)$ is the corresponding exploration-dependent safety margin. Since δ scales with ξ^2, larger probing amplitudes require a larger robustness buffer. In contrast, larger values of $\widehat{\eta}_h$ reduce this margin, thereby tightening the effective safety envelope in regions of rapid environmental variation.

Using (12), we define the enlarged safe set as the 0-superlevel set of $h_{\xi,L}$:

$$\mathcal{C}_{\xi,L} := \left\{x \in \mathbb{R}^n : h_{\xi,L}(x, \xi, \widehat{\eta}_h) \geq 0\right\}, \tag{13}$$

$$\partial\mathcal{C}_{\xi,L} := \left\{x \in \mathbb{R}^n : h_{\xi,L}(x, \xi, \widehat{\eta}_h) = 0\right\}, \tag{14}$$

$$\mathrm{Int}(\mathcal{C}_{\xi,L}) := \left\{x \in \mathbb{R}^n : h_{\xi,L}(x, \xi, \widehat{\eta}_h) > 0\right\}. \tag{15}$$

This construction is in the spirit of robust safety [3], but with a key extension: the enlargement is not fixed a priori and is instead adapted online using the local Lipschitz estimate extracted from exploration data.

Remark 3. It holds that $\mathcal{C} \subseteq \mathcal{C}_{\xi,L}$, and $\mathcal{C}_{\xi,L}$ expands monotonically with ξ. In the exploration-free case $\xi = 0$, one has $\delta(h(x), 0, \widehat{\eta}_h) = 0$, hence $h_{\xi,L}(x, 0, \widehat{\eta}_h) = h(x)$ and $\mathcal{C}_{\xi,L} = \mathcal{C}$. Hence, the proximity of $\mathcal{C}_{\xi,L}$ to the nominal safe set $\mathcal{C}$ is directly determined by the margin $\delta(h(x), \xi, \widehat{\eta}_h)$. □

Based on the enlarged set $\mathcal{C}_{\xi,L}$, we can now formalize the notion of Lipschitz-aware safe exploration.

Definition 1 (Lipschitz-Aware Input-to-State Safe Exploration (L-ISSf-Exp))**.** *The system* (7) *is said to undergo Lipschitz-aware input-to-state safe exploration with respect to the set $\mathcal{C}$ if, for every probing signal $e_u(\cdot)$ satisfying $\|e_u\|_\infty \leq \xi$, the set $\mathcal{C}_{\xi,L}$ defined by* (13)–(15) *is forward invariant.* □

Thus, when $\mathcal{C}_{\xi,L}$ remains forward invariant under exploration, the nominal set $\mathcal{C}$ is protected through a Lipschitz-aware robust buffer. In this case, we refer to $\mathcal{C}$ as an L-ISSf-Exp set. The next step is to characterize the control inputs that guarantee this property.

Definition 2 (Lipschitz-aware ISSf-Exp Control Barrier Function)**.** *Let $h : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, and suppose $\nabla h(x) \neq 0$ whenever $x \in \partial\mathcal{C}_{\xi,L}$. Then h is called a Lipschitz-aware ISSf-Exp-CBF for the system* (7) *if there exist constants $\kappa, \epsilon_0, k, \lambda, \eta_0, h_s > 0$ such that, for all $x \in \Omega$,*

$$\sup_{u \in \mathbb{R}^m} \Big[L_f h(x) + L_g h(x) u\Big] \geq -\kappa h(x) + \frac{\|L_g h(x)\|_2^2}{\bar{\epsilon}(h(x), \widehat{\eta}_h)}, \tag{16}$$

where $\bar{\epsilon}(h(x), \widehat{\eta}_h)$ is defined in (10)*.* □

The inequality (16) is written in terms of the corrective control input u only. The effect of the probing signal $e_u(t)$ is handled separately through the bound $\|e_u\|_\infty \leq \xi$, which enters the analysis via the robust margin $\delta(h(x), \xi, \widehat{\eta}_h)$. In this way, the barrier condition retains the standard optimization structure while still accounting for exploration-induced disturbances.

Remark 4. The conservatism of the exploration controller is governed by $\delta(h(x), \xi, \widehat{\eta}_h)$. For fixed $\widehat{\eta}_h$, this margin increases with $h(x)$, allowing less conservative exploration inside the safe set, while it shrinks as $h(x) \to 0$, yielding more cautious behavior near the boundary. For fixed $h(x)$, an increase in $\widehat{\eta}_h$ reduces $\bar{\epsilon}$ and hence δ, pulling $\mathcal{C}_{\xi,L}$ closer to $\mathcal{C}$. Thus, larger local Lipschitz estimates lead to more conservative control actions. □

Next, we show that if the control input satisfies (16), then the set $\mathcal{C}_{\xi,L}$ is forward invariant, thereby ensuring safe exploration of (7).

Theorem 1. *Consider the system under exploration* (7) *with $\|e_u\|_\infty \leq \xi$. Assume that $\widehat{\eta}_h(t)$ is updated at discrete sampling instants and held constant between updates, and let the applied corrective control input be chosen so that the inequality in* (16) *holds for all $x \in \Omega$. Then the set $\mathcal{C}_{\xi,L}$ defined in* (13)–(15) *is forward invariant. Consequently, the system* (7) *undergoes L-ISSf-Exp with respect to $\mathcal{C}$.*

Proof: Fix an inter-sampling interval $[t_k, t_{k+1})$ on which $\widehat{\eta}_h$ is constant. For any control input satisfying (16), the closed-loop dynamics (7) yield

$$\begin{aligned}\dot{h}(x,t) &= L_f h(x) + L_g h(x) u + L_g h(x) e_u(t) \\ &\geq -\kappa h(x) + \frac{\|L_g h(x)\|_2^2}{\bar{\epsilon}(h(x),\widehat{\eta}_h)} + L_g h(x) e_u(t).\end{aligned} \tag{17}$$

Applying Young's inequality,

$$a^\top b \geq -\frac{\|a\|_2^2}{\bar{\epsilon}} - \frac{\bar{\epsilon}}{4}\|b\|_2^2,$$

with $a = L_g h(x)^\top$, $b = e_u(t)$, and $\bar{\epsilon} = \bar{\epsilon}(h(x),\widehat{\eta}_h)$, gives

$$\dot{h}(x,t) \geq -\kappa h(x) - \bar{\epsilon}(h(x),\widehat{\eta}_h)\frac{\|e_u(t)\|_2^2}{4} \tag{18}$$

$$\geq -\kappa h(x) - \delta(h(x),\xi,\widehat{\eta}_h). \tag{19}$$

Since $h_{\xi,L}(x) = h(x) + \delta(h(x),\xi,\widehat{\eta}_h)/\kappa$, its derivative on $[t_k, t_{k+1})$ is

$$\dot{h}_{\xi,L}(x,t) = \left(1 + \frac{1}{\kappa}\frac{\partial \delta}{\partial h}(h(x),\xi,\widehat{\eta}_h)\right)\dot{h}(x,t). \tag{20}$$

Because

$$\frac{\partial \delta}{\partial h}(h(x),\xi,\widehat{\eta}_h) = \frac{k\xi^2}{4\big(1+\lambda\ln(\widetilde{\eta}_h/\eta_0)\big)} > 0,$$

the multiplier in (20) is strictly positive. On the boundary $x \in \partial\mathcal{C}_{\xi,L}$, one has $h_{\xi,L}(x) = 0$, i.e.

$$h(x) = -\frac{1}{\kappa}\delta(h(x),\xi,\widehat{\eta}_h).$$

Substituting this into (18) yields $\dot{h}(x,t) \geq 0$, and therefore $\dot{h}_{\xi,L}(x,t) \geq 0$ on $\partial\mathcal{C}_{\xi,L}$. Moreover,

$$\frac{\partial h_{\xi,L}}{\partial x}(x) = \left(1 + \frac{1}{\kappa}\frac{\partial \delta}{\partial h}(h(x),\xi,\widehat{\eta}_h)\right)\frac{\partial h}{\partial x}(x) \neq 0$$

whenever $h_{\xi,L}(x,\xi,\widehat{\eta}_h) = 0$. By Nagumo's theorem [19], $\mathcal{C}_{\xi,L}$ is forward invariant. Hence the system undergoes L-ISSf-Exp with respect to $\mathcal{C}$. ■

Finally, a QP-based optimization is proposed to synthesize a control law that ensures both stabilization and safety of the system (7) during exploration.

C. Lipschitz-aware Robust QP Controller (L-ISSf-QP)

We synthesize the exploration-time safety filter via the following robust QP. Given an initial admissible policy $u_0(x)$, the probing signal $e_u(t)$ is treated as a bounded matched disturbance, while the corrective term u_{cbf} is computed from

$$\begin{aligned} u_{\text{cbf}}^\star(x) = \arg &\min_{u_{\text{cbf}} \in \mathbb{R}^m} \ \frac{1}{2}u_{\text{cbf}}^\top u_{\text{cbf}} \\ \text{s.t.} \quad & L_f h(x) + L_g h(x)\big(u_0(x) + u_{\text{cbf}}\big) \\ & \geq -\kappa h(x) + \frac{\|L_g h(x)\|_2^2}{\bar{\epsilon}(h(x),\widehat{\eta}_h)}. \end{aligned} \tag{L-ISSf-QP}$$

The control applied to the plant during exploration is

$$u_{\text{safe}}(t) = u_0(x(t)) + e_u(t) + u_{\text{cbf}}^\star(x(t)).$$

Unlike standard ISSf-CBF formulations [3, 4], the proposed construction adapts the robustness gain online through $\widehat{\eta}_h$, yielding less conservative behavior in slowly varying regions and more conservative action when abrupt variations are detected. The method assumes that a nominal continuously differentiable safety function $h(x)$ is available a priori; its role is to robustify this nominal safety description against local perturbations, not to infer the safe set from data. Actuator bounds may be appended to (L-ISSf-QP), and a standard slack-variable relaxation can be used if guaranteed numerical feasibility is required. Theorem 1 pertains to the ideal hard-constrained case.

V. Safe Learning of Optimal Control Law

To join the proposed exploration stage with learning, we adopt the safety-aware policy-iteration framework of [12]. Starting from the admissible policy u_0 generated by the exploration stage, the learning phase minimizes the safety-aware cost

$$r_{\text{safe}}(x,u) = x^\top Q x + u^\top R u + B_\gamma(h(x)), \tag{21}$$

with corresponding performance index

$$V_{\text{safe}}(x_0,u) = \int_0^\infty r_{\text{safe}}(x(\tau),u(\tau))\,\mathrm{d}\tau. \tag{22}$$

The policy evaluation and improvement steps follow the standard safety-aware HJB recursion

$$\nabla W^{(i)\top}(x)\big(f(x) + g(x)u^{(i)}(x)\big) + r_{\text{safe}}(x,u^{(i)}(x)) = 0, \tag{23}$$

$$u^{(i+1)}(x) = -\frac{1}{2}R^{-1}g^\top(x)\nabla W^{(i)}(x). \tag{24}$$

Under the standard assumptions of [11, 12], the iterates remain safe and converge to a locally optimal safe policy. Since the learning stage is not the main novelty of this paper, and in the interest of space, we do not detail the safe learning aspect and refer the readers to Section 3 in [12] for the same.

The pseudo-algorithm is provided next.

VI. Simulations

Consider a nonlinear inverted pendulum system with dynamics, for all $t \geq 0$, given by

$$\begin{bmatrix}\dot{x}_1 \\ \dot{x}_2\end{bmatrix} = \begin{bmatrix} x_2 \\ \dfrac{3g}{2l}\sin x_1\end{bmatrix} + \begin{bmatrix} 0 \\ \dfrac{3}{ml^2}\end{bmatrix}u, \tag{25}$$

where x_1 is the pendulum angle, x_2 is the angular velocity, and u is the applied torque. The parameters are $m = 1\,\text{kg}$, $g = 9.81\,\text{m/s}^2$, and $l = 1\,\text{m}$. The initial stabilizing policy is computed using a CLF-QP and used to initialize the off-policy algorithm.

Exploration phase: A high-amplitude, jittery probing signal is used: $e_u(t) = \sum_{r=1}^{13}\xi\,\zeta_r\sin(\vartheta_r t)$, where $\vartheta = [1,3,7,11,13,15,17,19,21,23,25,27,29]$, each ζ_r is drawn from a zero-mean Gaussian distribution $\mathcal{N}(0,\sigma^2)$ truncated to $[-3\sigma,3\sigma]$, $\sigma = 0.15$, and $\xi = 10$. The exploration phase is conducted over $t \in [0,6.5]$ s, and state/input data are collected with sampling period $T_s = 0.01$ s. To assess the effectiveness of the proposed approach, we consider the angular-velocity safe set $h_{\min}(x) = x_2 + 5$ and $h_{\max}(x) =$

Algorithm 1 Lipschitz-Aware Safe Reinforcement Learning

```
Initialization: Choose $\epsilon_0, k, \lambda, \eta_0, h_s, \beta_h > 0$, buffer size $N_h$, noise bound $\xi$, initial state $x(0)$, and sampling time $T_s$.
Phase 1: Admissible initialization.
Compute a stabilizing admissible policy $u_0(x)$ using a CLF-QP.
Phase 2: Safe exploration.
Initialize $D_h \leftarrow \emptyset$.
For each sampling instant $t_k$
    Generate probing signal $e_u(t_k)$ with $\|e_u\|_\infty \leq \xi$.
    Measure $x(t_k)$ and evaluate $h(x(t_k))$.
    Update the moving buffer $D_h(t_k)$ with $(x(t_k), h(x(t_k)))$, keeping $|D_h(t_k)| \leq N_h$.
    Compute $\widehat{\eta}_h(t_k)$ from (8)–(9).
    Compute $\bar{\epsilon}(h(x(t_k)), \widehat{\eta}_h(t_k))$ and $\delta(h(x(t_k)), \xi, \widehat{\eta}_h(t_k))$.
    Solve (L-ISSf-QP) to obtain $u^\star_{\text{cbf}}(x(t_k))$.
    Apply
        $$u(t_k) = u_0(x(t_k)) + e_u(t_k) + u^\star_{\text{cbf}}(x(t_k)).$$
End For
Phase 3: Safe learning.
Run Safe Policy Iteration (Algorithm 2 in[12] ) initialized with $u_0$.
```

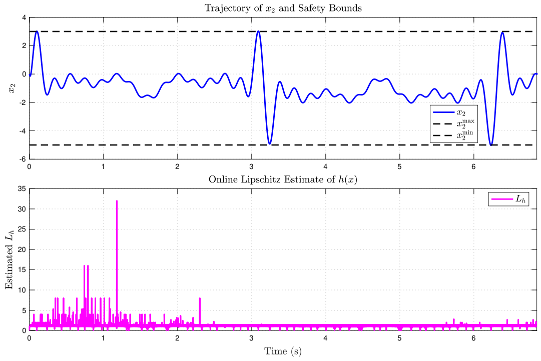

Fig. 1: Evolution of x_2 during exploration under static safety limits without Lipschitz-aware QP controller.

$3 - x_2$ so that $\mathcal{C} = \{x \in \mathbb{R}^2 : h_{\min}(x) \geq 0,\ h_{\max}(x) \geq 0\}$. We first study static limits (Fig. 1) and then abruptly varying limits (Fig. 2). *Case I: Fixed-margin Lipschitz-unaware baseline.* In this baseline, $\widehat{\eta}_h$ is not fed back to the controller, so $\bar{\epsilon}$ is kept constant and the QP cannot adapt to abrupt environmental changes. The controller is tuned with $\kappa = 35$. Under static safety limits, the state remains safe despite the probing signal (Fig. 1). When the safety limits are abruptly modified (Fig. 2), the fixed-margin controller is no longer sufficiently adaptive and safety violations may occur, even though the corresponding variations are reflected in the estimate $\widehat{\eta}_h$.

Case II: Exploration with Lipschitz adaptation. Under varying safety limits, $\widehat{\eta}_h$ is computed online using the KDE procedure of Section IV, with buffer size $N_h = 3$ and density threshold $\beta_h = 0.05$. The parameters are chosen empirically as $\epsilon_0 = 10^{-5}$, $k = 10^{-3}$, $\eta_0 = 10^{-3}$, and $\lambda = 10^5$, while $h_s > 0$ is selected so that $h(x) + h_s > 0$ over the operating domain. This tuning was found to provide sufficient sensitivity to abrupt changes in $\widehat{\eta}_h$, though it may vary across systems and probing signals.

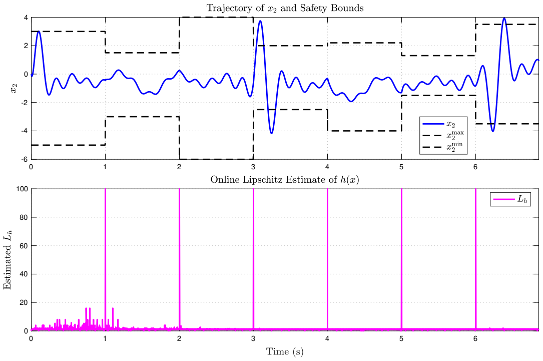

Fig. 2: Evolution of x_2 during exploration phase under abruptly varying safety limits without Lipschitz-aware QP controller.

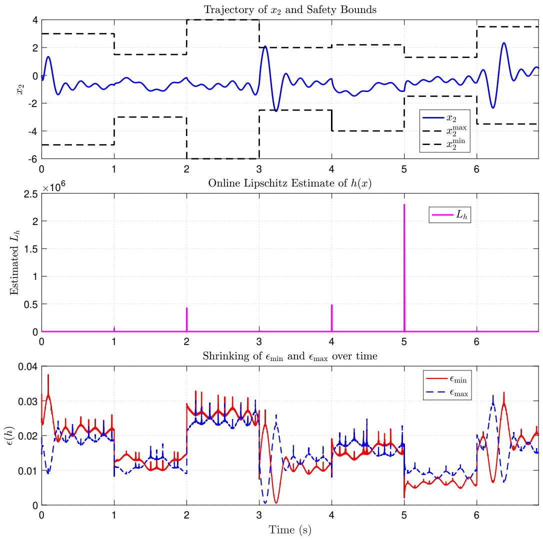

Fig. 3: Evolution of x_2 during exploration phase under abruptly varying safety limits under Lipschitz-aware QP controller.

Fig. 3 shows the state x_2 under the proposed L-ISSf-Exp-CBF controller, which prevents safety violations during exploration. Sharp changes in $\widehat{\eta}_h$ coincide with abrupt shifts in the safety limits or in the trajectory of x_2, supporting the central hypothesis of the paper: local Lipschitz estimates provide a practical indicator of abrupt environmental variations. We also plot the robustness gains $\bar{\epsilon}_{\min}$ and $\bar{\epsilon}_{\max}$, computed from $h_{\min}(x) = x_2 + 5$ and $h_{\max}(x) = 3 - x_2$, respectively. As expected, these quantities decrease as $\widehat{\eta}_h$ increases, thereby shrinking the effective set enlargement and making the controller more conservative when the estimated safety landscape becomes more variable.

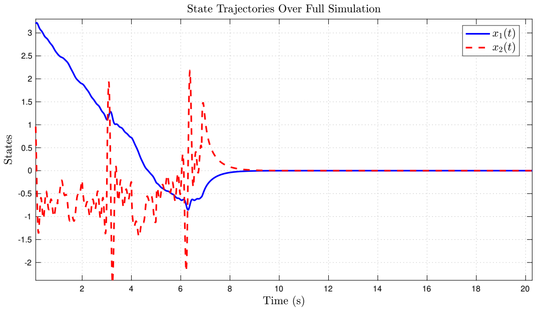

Fig. 4: Evolution of x_2 during the exploration as well as exploitation (learning) phase under the learned policy.

1) Exploitation (Learning) Phase: The reward function (21) uses $Q = \mathrm{diag}([0.1, 0.01])$, $R = 0.001$, and the RCBF penalty term $B_\gamma(h(x)) = -\log\left(\frac{\gamma h(x)}{\gamma h(x)+1}\right)$ with $\gamma = 0.7$, controlling the decay rate as the system moves away from the safety boundary. The critic NN $\phi(x)$ and actor NN $\psi(x)$ (activation functions) are chosen using polynomial basis functions respectively, as:

$$\begin{aligned}\phi(x) = \Big[&x_1^2, x_2^2, x_1x_2, x_1^4x_2^4, x_1^3x_2^2, x_1^2x_2^2, x_1x_2^3, x_1^6, x_2^6,\\ &x_1^5x_2, x_1^4x_2^2, x_1^3x_2^3, x_1^2x_2^4, x_1x_2^5, x_1^8x_2^2, x_1^7x_2, x_1^6x_2^2, x_1^4x_2^4,\\ &x_1^3x_2^5, x_1^2x_2^6, x_1x_2^7, x_1^{10}x_2^{10}\Big]^\top,\end{aligned} \tag{26}$$

$$\psi(x) = \Big[x_1, x_2, x_1x_2, x_1^2x_2^2, x_1^3, x_2^3, x_1^2x_2^3, x_1^4, x_2^4\Big]^\top. \tag{27}$$

The convergence of the weights of the actor and critic is observed in approximately 5 iterations. Fig. 4 shows the state evolution during both exploration and learning, with the safety constraint on x_2 preserved throughout.

VII. Conclusion

We propose a Lipschitz-aware safe exploration framework for model-based RL that captures unknown, abrupt environmental variations through local Lipschitz estimates. By leveraging a KDE approach to estimate the local Lipschitz constant of the safety function, we can detect sudden shifts in safety limits and state dynamics. The proposed Lipschitz-aware Input-to-State Safety Exploration scheme incorporates these Lipschitz estimates into a robust QP-based controller under Lipschitz-aware CBF constraints, ensuring no safety violations during exploration. This guarantees forward invariance of an enlarged safe set even amid aggressive exploration noise and abrupt unmodeled environmental changes. By integrating safe exploration with safe learning, this framework represents the first steps towards a comprehensive safe RL scheme capable of operating in rapidly changing, unknown nonlinear environments. Future work will investigate event-triggered mechanisms for safe exploration.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Mayank Shekhar Jha, Didier Theilliol, and Philippe Weber. "Model-free optimal tracking over finite horizon using adaptive dynamic programming". In: *Optimal Control Applications and Methods* 44.6 (2023), pp. 3114–3138.

[8] Satya Marthi, Mayank S Jha, Soha Kanso, Jean-Christophe Ponsart, and Didier Theilliol. "On-policy Safe Reinforcement Learning under Input Saturation and State Constraints for Nonlinear Discrete Time Systems". In: *2025 IEEE 64th Conference on Decision and Control (CDC)*. IEEE. 2025, pp. 4817–4824.

[9] Théo Rutschke, Mayank Shekhar Jha, and Hugues Garnier. "Neural Ordinary Differential Equations based System Identification for Reinforcement Learning with Provable Guarantees". In: *2025 IEEE 64th Conference on Decision and Control (CDC)*. IEEE. 2025, pp. 3848–3855.

[10] 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.

[11] 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.

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

[14] 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.

[15] Si Wu, Tengfei Liu, Magnus Egerstedt, and Zhong-Ping Jiang. "Quadratic programming for continuous control of safety-critical multiagent systems under uncertainty". In: *IEEE Transactions on Automatic Control* 68.11 (2023), pp. 6664–6679.

[16] Ankush Chakrabarty, Devesh K Jha, Gregery T Buzzard, Yebin Wang, and Kyriakos G Vamvoudakis. "Safe approximate dynamic programming via kernelized lipschitz estimation". In: *IEEE transactions on neural networks and learning systems* 32.1 (2020), pp. 405–419.

[17] Soha Kanso, Mayank Shekhar Jha, and Didier Theilliol. "Safe Reinforcement Learning Tracking Control based on Tunable Input-to-State Safe Control Barrier Function". In: *2025 American Control Conference (ACC)*. IEEE. 2025, pp. 3103–3108.

[18] Soha Kanso, Mayank Shekhar Jha, and Didier Theilliol. "Reinforcement learning-based degradation-tolerant control design for affine nonlinear systems". In: *Asian Journal of Control* (2025).

[19] 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).