2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

Learning Adaptive Safety for Multi-Agent Systems

Luigi Berducci[1a], Shuo Yan

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10611037

***Abstract*— Ensuring safety in dynamic multi-agent syste challenging due to limited information about the other ag Control Barrier Functions (CBFs) are showing promis safety assurance but current methods make strong assump about other agents and often rely on manual tuning to ba safety, feasibility, and performance. In this work, we delv the problem of adaptive safe learning for multi-agent sy with CBF. We show how emergent behaviour can be profou influenced by the CBF configuration, highlighting the nec for a responsive and dynamic approach to CBF design present ASRL, a novel adaptive safe RL framework, to automate the optimization of policy and CBF coefficients, hance safety and long-term performance through reinforce learning. By directly interacting with the other agents, ASRL learns to cope with diverse agent behaviours and maintains the cost violations below a desired limit. We evaluate ASRL in a multi-robot system and competitive multi-agent racing, against learning-based and control-theoretic approaches. We empirically demonstrate the efficacy of ASRL, and assess generalization and scalability to out-of-distribution scenarios.**

I. Introduction

Safety is an prominent concern in the design of learning algorithms, especially for safety-critical applications. Control barrier functions (CBFs) have emerged in this context as a very powerful formal approach to ensuring safety [1]–[3]. Moreover, the integration of CBFs in reinforcement learning (RL) holds a huge potential for safe exploration [4]–[8].

However, the success of CBFs in RL is often confined to simple settings, with single or multiple cooperative agents and very limited interactions. This is because the intricate interplay among many agents poses unique challenges to the design of the CBFs and their associated *(extended) class-$\mathcal{K}$ functions, which are continuous stricly-increasing functions* $\gamma : \mathbb{R} \to \mathbb{R}$ *such that* $\gamma(0) = 0$. These functions control the rate with which the agent can approach the safe-set boundary.

While manual tuning of the class-$\mathcal{K}$ functions is feasible in simple tasks, it becomes challenging in multi-agent environments due to the unpredictable effects of small changes in their parameters. The richness of interactions and limited information about other agents' policies make it difficult to trade-off safety and long-term performance by adjusting just a few parameters. Previous work on Adaptive CBFs have primarily concentrated on enhancing the feasibility of the optimization problem, by introducing time-varying coefficients within the CBF condition [9]. However, it's worth noting that these approaches have been predominantly applied in single-agent or cooperative environments, often relying on the availability of substantial historical data for optimization [10], [11]. These methods face two key challenges:

- *Overlooking long-term objectives*: the narrow focus on feasibility and short-term performance within the adaptive CBF framework, struggles in capturing long-term objectives and potentially leads to sub-par solutions.
- *Prior-data scarcity*: the assumption that historical data is available does not always hold, especially in non-cooperative settings where other agents may be reluctant to reveal their strategies, thereby hindering the achievement of sufficient coverage of diverse strategies.

To address these two challenges in interactive multi-agent environments, we propose a novel approach based on RL and adaptive CBFs. In order to account for the lack of knowledge with regard to the other agents' strategies, we exploit direct interactions with these agents, to uncover their intentions. Our **main contributions** in this paper are the following:

1) *An adaptive safe-RL framework (ASRL)*, where a low-level CBF-controller ensures safety and a high-level one optimizes policy and state-dependent CBF coefficients.
2) *A model-free learning approach*, which is based on RL for efficient adaptation to different agents and scenarios through direct interaction with these agents.
3) *A comprehensive evaluation of ASRL* in multi-agent environments, in order to assess the adaptation to different types of agents and degrees of cooperation.

As shown in Figure 1, by combining a model-based low-level control layer with model-free RL, ASRL enhances adaptiveness to diverse behaviors exhibited by other agents, relieving the engineers from the burden of manual tuning the CBF, in favour of a systematic approach to optimize general long-term objectives and trade-off safety and performance.

Motivating Example: Consider a navigation task where multiple robots have distinct starting positions and specific

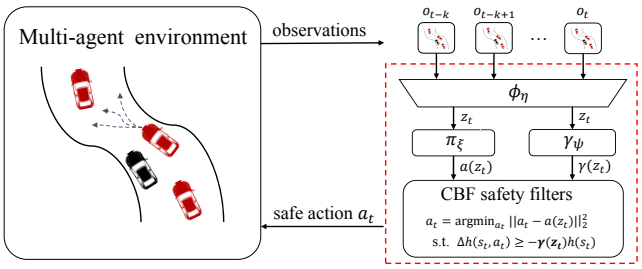

Fig. 1: The proposed hierarchical adaptive framework for multi-agent systems, where a policy π_ξ and a safety module γ_ψ are jointly optimized for safe and adaptive interaction.

[1]Institute of Computer Engineering, TU Wien
[2]Dept. of Electrical and Systems Engineering, University of Pennsylvania
[a]Correspondance to: luigi.berducci@tuwien.ac.at
Code and supplementary material at https://github.com/luigiberducci/learning_adaptive_safety

goals. The controlled *ego* robot needs to reach its goal while avoiding collisions with other robots, without any knowledge of their parameterization. To ensure safe navigation, we equip the ego robot with a CBF, acting as a protective safety shield. However, the emergent behavior of the ego robot can vary significantly by adjusting the coefficient γ of the CBF class-$\mathcal{K}$ function, as shown in Figure 2 (left). In some simulations, the ego fails to reach its goal due to cautious maneuvers dictated by the CBF condition, while the same controller successfully completes the task in other configurations.

Two factors contribute to this: (1) the ego robot adapts its maneuvers based on the CBF condition, varying assertiveness levels; (2) other robots react to the ego's actions, leading to configurations that can aid or hinder task completion. Figure 2 (right) supports this hypothesis by showing how diverse CBF coefficients influence the long-term performance of the ego agent, measured by success and collision rates. The optimal coefficients depend on scenario-specific characteristics, such as the number of agents and their parameters. This underscores the importance of an adaptive approach, which will be detailed in the following sections of this work.

II. Background

Consider the stochastic game $(\mathcal{I}, S, \bar{A}, f, r, \rho_0, T, \alpha)$, where $\mathcal{I} = \{1, 2, \cdots, q\}$ denotes the set of q agents, S and $\bar{A}$ are the set of states and joint actions, $f : S \times A \to S$ is the deterministic transition function, $r : S \times A \times S \to \mathbb{R}$ is the reward function, ρ_0 represents the distribution of initial conditions, $T \in \mathbb{N}$ denotes the time horizon, $\alpha \in [0, 1]$ is the discount factor, to avoid confusion with the γ adopted in CBF. At timestep t, each agent i picks an action a_t^i according to its policy π_i and the system state evolves according to the joint action $a_t = \times_{i \in \mathcal{I}} a_t^i$ by discrete-time dynamics

$$s_{t+1} = f(s_t, a_t). \tag{1}$$

We consider the problem of finding an optimal policy for the *ego* agent π_1, assuming the non-controlled agents $\pi_i, i > 1$, have unobservable parameters distributed according to ρ_0.

In the following, we denote the ego with π and formulate the problem as a constrained partially observable Markov decision process (CPOMDP) $(S, A, \Omega, O, f, r, h, \rho_0, T, \alpha)$ where the actions A refer to π, the observations Ω consists of the observable states without the other agents' parameters obtained by $O : S \to \Omega$, h is a continuously differentiable function delimiting the safe set of states for the ego agent. We define the safe set, $\mathcal{C}$, by the superlevel set of h:

$$\mathcal{C} = \{s \in S : h(s) \geq 0\}. \tag{2}$$

Definition 1. (Forward invariance and safety) The set $\mathcal{C}$ is *forward invariant* if for every $s_0 \in \mathcal{C}$, $s_t \in \mathcal{C}$ holds for all t. If $\mathcal{C}$ is forward invariant, we say the system (1) is safe.

Definition 2. (CBF [12]) Given a set $\mathcal{C} \subset \mathbb{R}^n$ defined by (2), the continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ is a *discrete-time control barrier function* (CBF) for dynamical system (1) if there exists $\gamma \in [0, 1]$ such that for all $s_t \in \mathcal{C}$,

$$\sup_{a_t \in A} \Big[h\Big(f(s_t, a_t)\Big) + (\gamma - 1) h(s_t) \Big] \geq 0. \tag{3}$$

Note that the parameter γ influences the conservativness of agent's behaviour: it will be less conservative (i.e., approaching the safe boundary) as γ goes to 1. However, γ is fixed in the above vanilla CBF definition, which implies the fixed degree of conservativness. To overcome this limitation, we introduce the adaptive version of discrete-time CBF.

Definition 3. (Adaptive control barrier function) Given a set $\mathcal{C} \subset \mathbb{R}^n$ defined by (2), the continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ is a *discrete-time adaptive control barrier function* (Adaptive CBF) for dynamical system (1) if for all $s_t \in \mathcal{C}$, there exists $\gamma(s_t) \in [0, 1]$ such that

$$\sup_{a_t \in A} \Big[h\Big(f(s_t, a_t)\Big) + (\gamma(s_t) - 1) h(s_t) \Big] \geq 0. \tag{4}$$

The Adaptive CBF differs from (3) by the state-dependent function $\gamma : S \to [0, 1]$. We now demonstrate that state-dependent coefficients do not hinder the safety guarantees of CBF. For any potentially unsafe nominal action a_t^{nom}, we can obtain a safe action solving the optimization problem:

$$\begin{aligned} a_t = \ & \underset{a_t \in A}{\text{argmin}} \quad \|a_t - a_t^{nom}\|_2^2 \\ & \text{s.t.} \qquad h(f(s_t, a_t)) + (\gamma(s_t) - 1) h(s_t) \geq 0. \end{aligned} \tag{5}$$

Lemma 1. *For dynamical system (1), if the problem (5) is feasible for all $s \in \mathcal{C}$, then the derived controller from (5) renders set $\mathcal{C}$ forward invariant, i.e., safety is preserved.*

Proof. For any initial state $s_0 \in \mathcal{C}$, we can derive that:

$$\begin{aligned} h(s_t) &= h(f(s_{t-1}, a_{t-1})) \geq (1 - \gamma(s_{t-1}))\, h(s_{t-1}) \\ &\geq (1 - \gamma(s_{t-1}))\,(1 - \gamma(s_{t-2}))\cdot h(s_{t-2}) \\ &\cdots \\ &\geq \prod_{i=0}^{t-1} (1 - \gamma(s_i))\, h(s_0) \geq 0 \end{aligned} \tag{6}$$

which implies that safety is preserved at any time t. □

Practical CBF usability in multi-agent systems comes with a few important remarks. Problem (5) is tipically non-convex, and prior work focused on linear CBFs and convex formulations, for improved efficiency and optimal solutions [4].

In contrast, ASRL does not assume the ability to optimally solve (5). Instead, ASRL adaptiveness uses class-$\mathcal{K}$ functions, allowing users to provide any CBF. Moreover, handling multiple CBF constraints and input limits introduces feasibility issues [13], [14]. This is further exacerbated in multi-agent systems, where limited information of other agents preclude full safety guarantees. These uncertainties can be quantified and integrated *explicitly* in robust CBF formulations, with probabilistic guarantees [15]. In this work, the uncertainty is *implicitly* learned in the model, the feasibility ensured by introducing slack variables, and the safety relaxed into a chance constraint, as formulated in next section.

III. Problem Statement

We consider the problem of learning a policy π_θ with parameters θ for an agent operating in a multi-agent environment, with partial observability of the other agents' parameters. We formulate it as a CPOMDP with a cost function h

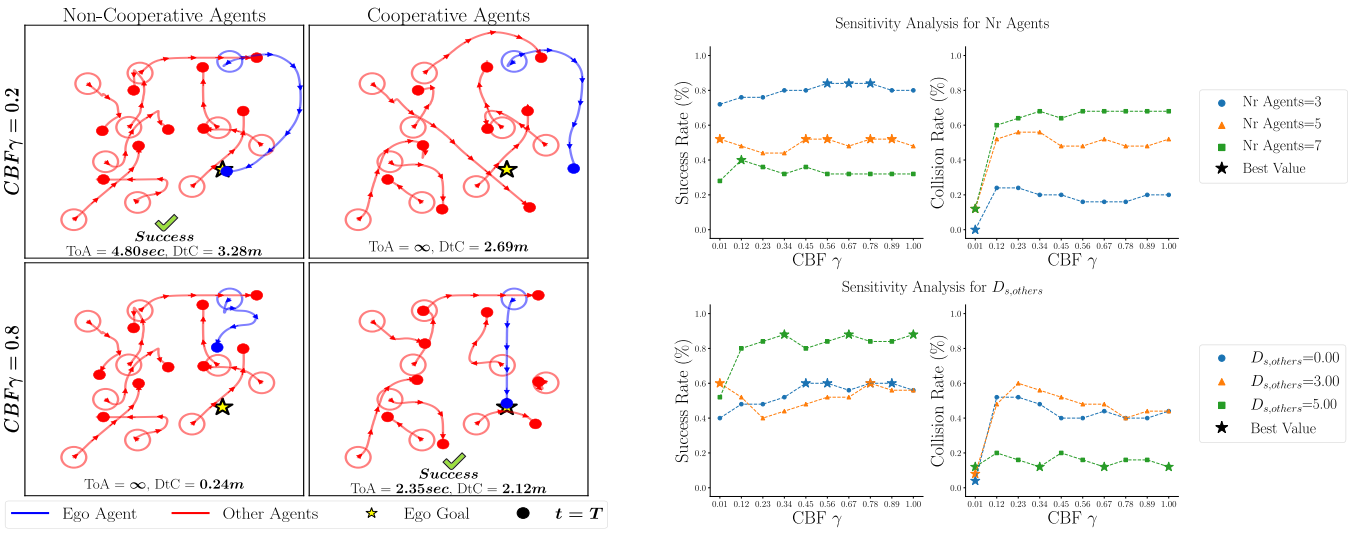

Fig. 2: **Left:** Simulations with non-cooperative (left) and cooperative agents (right) under different CBF coefficients γ. Time of Arrival (*ToA*) and Minimum Distance to Collision (*DtC*) are reported for the ego agent (*blue*). The long-term effects diverge based on the coefficients, showing the importance of adaptation. **Right:** Average performance of CBF coefficients under different number of agents (*top*) and safety distances $D_{s,others}$ (*bottom*) of other agents' controllers ($n = 25$).

and aim to find a solution that keeps the occurrence of safety violations below a desired level of tolerance $d \in [0, 1]$.

Formally, the policy optimization problem is defined as:

$$\max_{\theta} \mathcal{J}_R(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\big[\sum_{t}^{H} \alpha^t\, r(s_t, a_t, s_{t+1})\big] \qquad (7)$$
$$\text{s.t. } \mathcal{J}_C(\theta) = \mathbb{P}_{\tau\sim\pi_\theta}(\min_t h(s_t) < 0) \le d$$

where the trajectory $\tau = (s_0, a_0, \cdots, s_H)$ results from the interaction of π with the agents $\pi_{i>1}$, given the initial distribution $s_0, \pi_{i>1} \sim \rho_0$, and dynamics $s_{t+1} = f(s_t, a_t)$.

IV. Adaptive Safe Reinforcement Learning

We present ASRL, the main contribution of this work, an adaptive framework for multi-agent systems, which combines low-level model-based control and model-free RL in a hierarchical fashion, and the associated optimization algorithm.

Hierarchical Model Architecture. We structure the agent π into a high-level model, which selects the desired goal and provides an adaptive class-$\mathcal{K}$ function, and a low-level layer, which enforces the system safety using the barrier function h, the actions, and the coefficients from the high-level model.

To address partial observability of other agents, we adopt a novel multi-head actor with the following components:

$$\text{Representation model:} \quad z_t = \phi_\eta(o_{t-k:t}) \qquad (8)$$
$$\text{Policy model:} \quad a_t \sim \pi_\xi(\ \cdot\ |\ z_t) \qquad (9)$$
$$\text{Safety model:} \quad \gamma_t \sim \gamma_\psi(\ \cdot\ |\ z_t) \qquad (10)$$

It consists of a representation model ϕ_η which encodes the past k observations $o_{t-k:t}$ into an embedding z, a policy head π_ξ which produces action a_t, and an adaptive-safety head γ_ψ which outputs the CBF coefficient γ_t. Our multi-head model is constructed with a specific emphasis on modularity, thereby enforcing a separation of concerns in design. The joint training of these components is carried out as a single integrated model, with the parameters denoted as $\theta = (\eta, \xi, \psi)$, and the details described in the next section.

Learning Adaptive Behaviors. We solve the Optimization Problem (7), considering its unconstrained relaxation:

$$\min_{\lambda\ge 0}\max_{\pi_\theta\in\Pi} \mathcal{J}(\theta, \lambda) = \min_{\lambda\ge 0}\max_{\pi_\theta\in\Pi} \mathcal{J}_R(\theta) - \lambda\, \mathcal{J}_C(\theta) \qquad (11)$$

where $\mathcal{J}$ is the Lagrangian, and $\lambda \ge 0$ is the Lagrange multiplier which acts as penalty weight. The two optimizations steps are interleaved till convergence, seeking for a saddle point of the original problem which is a feasible solution.

Policy Update. Considering the model-free setting due to the lack of knowledge of other agents, the true return and cost distributions are induced by the policy rollouts and unknown. We use a policy-gradient algorithm [16], jointly optimizing an actor π_θ and a critic v_ζ models. The critic simply regresses the value estimates $v_{target}(z_t)$, minimizing the loss:

$$\mathcal{L}_v(\zeta) = \mathbb{E}_t\big[(v_\zeta(z_t) - v_{target}(z_t))^2\big] \qquad (12)$$

The actor is updated by maximizing the following loss

$$\mathcal{L}_\pi(\theta) = \mathcal{L}_R(\theta) - \lambda_k \mathcal{L}_C(\theta) + \beta \mathcal{L}_{ent}(\theta) \qquad (13)$$

where λ_k is the Lagrange multiplier introduced in Eq. (11) at the k-th update, $\mathcal{L}_R, \mathcal{L}_C$ denote the surrogate clipped losses for cumulative rewards and costs [16], and $\mathcal{L}_{ent}$ denotes the entropy bonus for exploration. We use generalized advantage estimation (GAE) [17] to trade-off bias and variance in the advantage estimates $\hat{A}_{R,t}, \hat{A}_{C,t}$ for return and cost respectively. The surrogate clipped losses are defined as:

$$\mathcal{L}_R(\theta) = \mathbb{E}_t\big[min(r_t(\theta)\hat{A}_{R,t}, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{R,t})\big]$$
$$\mathcal{L}_C(\theta) = \mathbb{E}_t\big[min(r_t(\theta)\hat{A}_{C,t}, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{C,t})\big]$$
$$r_t(\theta) = \frac{\pi_\theta(a_t|z_t)}{\pi_{\theta_{old}}(a_t|z_t)}$$

Lagrange Multiplier Update. The Lagrange multiplier plays as an adaptive penalty in the unconstrained problem, to make the infeasible solutions sub-optimal. We update the Lagrangian multiplier with the PID update rule [18], because of its effectiveness and simplicity in the implementation. This update rule resembles the tuning of PID controllers to correct oscillations and overshooting of traditional Lagrangian methods [19]. The update rule at iteration k is as follows:

$$\lambda_k \leftarrow (K_P\Delta_k + K_I I_k + K_D\delta_k)_+ \tag{14}$$

where $K_P, K_I, K_D \in \mathcal{R}_+$ are hyperparameters for the proportional, integral and derivative errors, defined as:

$$\Delta_k = \mathcal{J}_{C,k} - d \tag{15}$$

$$I_k = (I_{k-1} + \Delta_k)_+ \tag{16}$$

$$\delta_k = \mathcal{J}_{C,k} - \mathcal{J}_{C,k-1} \tag{17}$$

Low-level Control Design. This section presents the design of CBF for multi-agent systems, which involves two steps:

1) *State and dynamics identification*: we model the underlying system's dynamics. This step can follow first principles, employing physical laws and motion equations, data-driven approaches, or a combination of these.
2) *CBF design*: we design CBFs to enforce safety, mapping states to numerical values. This step presents several challenges in formalizing safety, defining barriers, efficient modeling and parameters selection to ensure feasibility while balancing safety and performance.

We exemplify this methodology in the motivating example.

Multi-robot system. Consider a system with n agents [20], where each agent i has control-affine dynamics:

$$s_{i,t+1} = \begin{bmatrix} p_{i,t+1} \\ v_{i,t+1} \end{bmatrix} = \begin{bmatrix} I & \Delta t \\ 0 & I \end{bmatrix} \begin{bmatrix} p_{i,t} \\ v_{i,t} \end{bmatrix} + \begin{bmatrix} 0 \\ \Delta t \end{bmatrix} a_{i,t} \tag{18}$$

where Δt is the discrete time-step, $p_i \in \mathcal{R}^2$, $v_i \in \mathcal{R}^2$, $a_i \in \mathcal{R}^2$ denote the position, velocity and acceleration of robot i respectively. We can write the joint multi-agent system as:

$$s_{t+1} = \begin{bmatrix} p_{t+1} \\ v_{t+1} \\ z_{t+1} \end{bmatrix} = \begin{bmatrix} f_p(s_t) \\ f_v(s_t) \\ f_z(s_t) \end{bmatrix} + \begin{bmatrix} g_p(s_t) \\ g_v(s_t) \\ g_z(s_t) \end{bmatrix} a \tag{19}$$

where $p \in \mathcal{R}^2$ and $v \in \mathcal{R}^2$ denote the position and velocity of the ego agent, $a \in \mathcal{R}^2$ denotes the ego action, and $z \in \mathcal{R}^{n-4}$ denotes the other agent's states. The real-valued functions f_o, g_o are known for $o \in p, v$. However the other agents' actions are unknown for the ego (i.e., $g_z = 0$). Without loss of generality, they can be assumed to be a function of the joint state s_t and part of the dynamics f_z. We consider the CBFs as pairwise safety constraints between the ego agent i and any other agent $j \neq i$:

$$h(x) = \frac{\Delta p_{ij}^T}{||\Delta p_{ij}||}\Delta v_{ij} + \sqrt{a_{max}(||\Delta p_{ij}|| - D_s)} \tag{20}$$

where $a_{\max}$ denotes the maximum braking that the ego agent can apply to avoid a collision, Δp_{ij} represents the relative position $p_i - p_j$, and Δv_{ij} the relative velocity $v_i - v_j$.

Multi-agent Racing. We consider a second use case of competitive multi-agent racing. The dynamics are modeled using an Euler-discretization of the kinematic bicycle model as in [21]. We consider two safety specifications for collision with walls and opponents, and model them using distance on Frenet and Cartesian coordinates, respectively. For conciseness, we describe the dynamics and CBFs in the Appendix.

V. Experiments

In this section, we describe the experiments to evaluate our adaptive safe-learning approach in multi-agent systems.

Simulation. We conducted our experiments in the multi-agent environments presented in the previous section. For the multi-robot system, we use the simulator and CBF from [15] with its simplest collision-avoidance formulation. For the multi-agent racing system, we use the F1tenth simulator [22], which provides simulation of multiple vehicles and sensory inputs. In both the environment, the CBF uses a constant-velocity model (CVM) for the other-agents behaviors.

Training. We implemented the ASRL algorithm with the `omnisafe` library [23]. During training, we randomize the starting conditions and collect episodes of 15 seconds. The agent observes the last 5 states and learns with progress-based reward and sparse cost signals: for the multi-robot system, the cumulative reward is 1 for reaching the goal location and the cost is 1 for collision with any opponent; for the racing system, the reward is proportional to the relative distance in front of other vehicles and the cost is 1 for collisions. We evaluate the agent by averaging the reward and cost with a moving average over the last 100 episodes and train the agents for 1 million steps. More details on the environments and training are reported in the Appendix.

Agents' Randomization: To create the condition for adaptiveness, we randomize the number of agents and policies in each environment. In the multi-robot system, we select between 3 and 7 agents at each episode and randomize their policies through the safety distance D_s used to avoid obstacles. In the multi-agent racing system, we simulate 2 vehicles starting in front of the ego vehicle and tracking a reference line with velocity profile randomly scaled by a factor normally distributed ($\mu = 0.78$, $\sigma = 0.05$).

Comparison End-to-End. In Figure 3, we compare ASRL with the following state-of-the-art safe RL baselines: PPO Lagrangian, DDPG Lagrangian and TD3 Lagrangian, which are safe versions of the on-policy PPO [16], and off-policy DDPG [24] and TD3 [25] respectively; CPO, a trust-region method with near-satisfaction guarantees [26]; IPO, an interior-point policy optimization method [27]; PPO Sauté, a state-augmentated PPO on the Sauté MDP [28].

In both environments, a noticeable performance gap emerges with on-policy algorithms which struggle in achieving safe solutions within the desired limit. Their on-policy nature may require extended training periods to attain safe and optimal results. Meanwhile, in the multi-robot environment, the off-policy algorithms DDPG-Lag and TD3-Lag achieve safe solutions but at the expense of poor overall

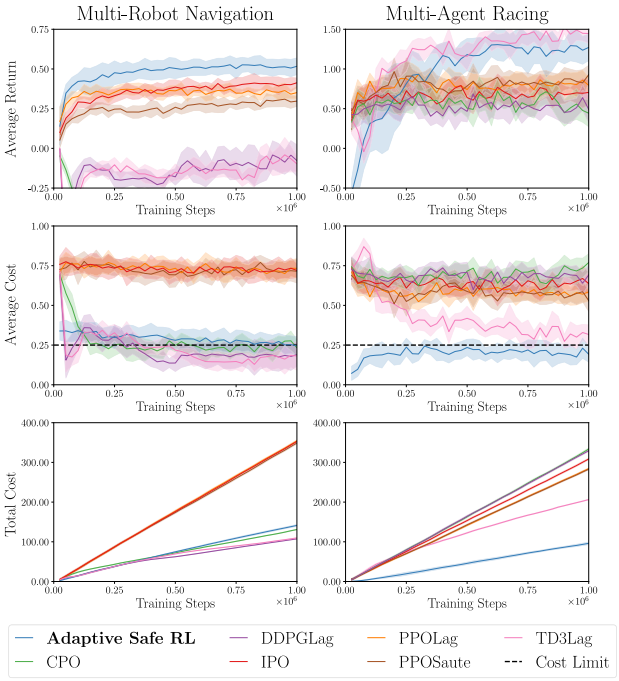

Fig. 3: Learning curves of our approach and safe-RL baselines. Return, cost, and total cost averaged over 3 runs.

performance. Similarly, in the multi-agent racing environment, on-policy algorithms continue to fail, while notably, TD3-Lag exhibits slow but consistent progress, reaching a near-optimal level of performance by the end of training.

In contrast, our ASRL approach, incorporating trainable policies and CBF coefficients, fastly converges to high returns while consistently staying around the desired cost limit. Notably, in the multi-robot task, it starts above the cost limit and gradually reaches it, whereas in the racing task, it initiates below the limit and gradually adjusts safety levels to approach it. Our results suggest that our integration of CBF into the agent model improves exploration during training, resulting in a reduction in cumulative violation costs comparable to or better than off-policy methods, all while achieving significantly higher performance (Fig. 3, bottom).

Ablation Study on Learning Components. In this experiment, we evaluate the impact of adaptive safety and demonstrate its domain-adaptation skills. To assess adaptive safety in isolation, we replace the policy module with a non-trainable controller. During training, we use a Perturbed Gaussian policy [29] to foster exploration. Details on training and controllers are reported in the supplementary material.

We evaluate the performance of the ablate model against traditional control-theoretic approaches: Standard CBF (S-CBF), which uses CBF with fixed class-$\mathcal{K}$ coefficients, and Optical-Decay Adaptive CBF (OD-CBF), which adapts the coefficients to ensure point-wise feasibility [12]. To account for the fact that these methods' performance are highly sensitive to the coefficients initialization, we discretize the range of CBF parameters in 10 values to cover most of the configurations. We collect 100 episodes for each configuration and report the performance in Figure 4. We observe comparable performance of OD-CBF and S-CBF in both the environments, confirming that optimal-decay adaptiveness might improve feasibility of the QP problem but cannot capture long-term objectives, as sparse and delayed events of success or collision. Conversely, our ablated model outperforms the baselines by simply adapting the CBF coefficients based on interactions with other agents. This strongly suggests that learning of adaptive safety can substantially enhance the performance of existing controllers.

Fig. 4: Comparison of the ablate model with control-theoretic approaches using the same nominal controller. The bars show the mean rate with min/max delimiters for the same method. Performance for trained models are averaged over 3 runs.

Generalization and Scalability. To evaluate the generalization capabilities of our trained agent in multi-agent systems, we consider diverse racing scenarios including: (1) varying the number of agents, (2) in-distribution planners with varying velocity profiles, and (3) out-of-distribution planners with new strategies and varying velocity profiles. We focus on in- and out-distributions opponents, deliberately excluding cross-track generalization because competitive high-speed racing demands specialized strategies, even to the extent of overfitting to track conditions. For each scenario, we sample 10 starting positions with the ego behind (Figure 5, top left) and run simulations for 60 seconds. We consider collision or lap completion as termination conditions and measure the final positioning (*rank*) as common in racing competitions.

As shown in Figure 5, the trained agent exhibits competitive performance, often reaching 1st and 2nd place despite the initial positional disadvantage. Training with 2 opponents proves sufficient for generalization to races with many agents. Up to 8 opponents, the agent consistently secures a podium spot with an average rank below 3rd place. However, with 9 agents or more, the average rank exceeds the 4-th position due to the limited time horizon for race completion.

For *in-distribution* planners, the agent shows good adaptation and remains robust to faster velocity profile. Notably, the rank increase appears directly linked to the training distribution. For *out-distribution* planners, we consider reactive (FTG [30]), waypoint follower (PP [31]), and sampling-based planners (Lattice [32]) with different velocities. Our agent outperforms FTG, as expected due to its reactive nature

which lacks of any global raceline. Moreover, we observe a competitive racing style against the Lattice planner, a robust baseline for comparison. Notably, our agent maintains a high performance even against high-velocity profiles (*right-most bars*), suggesting the ability to learn characteristics racing behaviors and reuse them against unseen opponents.

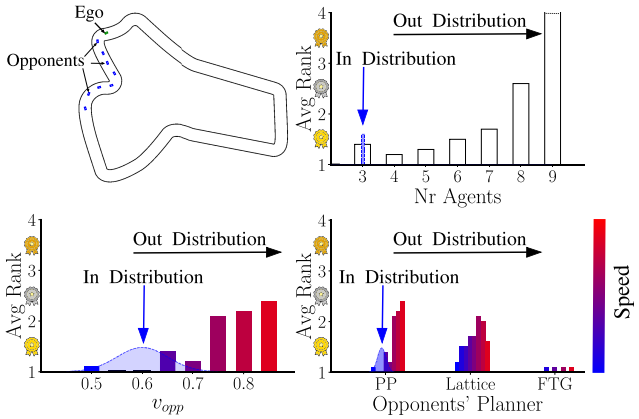

Fig. 5: Generalization in multi-agent racing (*top left*). Performance measure the ego rank (*lower is better*) under previously unseen number of agents, velocity profiles v_{opp}, and planners. The training distribution is overlayed (*blue*).

Summary of Results. We assess our adaptive approach in two multi-agent systems with a range of diverse agents. The experiments revealed that our hierarchical integration of CBF facilitated convergence to near-optimal agents, outperforming a variety of safe RL baselines. Moreover, our ablation study isolated the impact of Adaptive CBF, showcasing superior adaptiveness than traditional control-theoretic methods. Finally, empirical evidence demonstrated our agent's ability to adapt and generalize across various racing scenarios, including unseen opponents and high-velocity profiles.

VI. Related Work

Safe RL via CBF. Safe RL has drawn much attentions to prevent visiting unsafe states in safety-critical systems [33]–[36]. The application of CBF in Safe RL has been proposed in [4] and is getting popular because of its safety guarantees and computational efficiency [6], [37]. Existing works train an RL agent to propose actions and use vanilla CBF to enforce safety. However, to the best of our knowledge, we are the first jointly training state-dependent CBF coefficients with an RL policy to ensure bounded chance of violations.

Adaptive CBFs. Several works focus on improving feasibility and performance of CBF-controllers [9], [11], [12], [38], [39], mostly focusing on single-agent or cooperative systems. In these settings, the CBF coefficients are optimized through gradient-based methods [38] or policy distillation of a network with differentiable CBF layer, under the assumption of available expert demonstrations [11]. None of these works focus on multi-agent environments and adaptation to changing policies. In contrast, our approach leverages discrete-time CBF, which better fits MDP theory, to train a model tailored for multi-agent adaptation with online RL. CBF coefficients are updated in [39] based on the level of cooperation of other agents towards the ego. However, they assume the ego agent knows the other agents' actions information beforehand. Morevoer, they do not offer a direct way to update CBF coefficients but only mentioning that the derivative of them is monotonically increasing w.r.t. the level of cooperation, still requiring user intervention. In contrast, we do not rely on such an assumption and leverage RL to train the coefficients, thus replacing any manual effort with a systematic methodology. Moreover, we demonstrate our approach in a challenging multi-agent racing scenario.

Multi-agent CBF. In multi-agent systems, prior research proposed CBF for collision-free behavior [20], [40]. Among these, [41] proposes a scalable decentralized approach to control multiple agents, [15] presents a robust CBF with uncertainty model learned from data, and [10] focuses on the joint optimization of control policy and CBFs. However, these works do not operate within an RL setting and do not consider adaptation to other agents. Moreover, they mostly rely on fixed class-$\mathcal{K}$ function.

VII. Conclusions

We present Adaptive Safe RL (ASRL) for multi-agent systems with partial observability from interactions with other agents. We combine model-free RL and adaptive CBF to optimize long-term objectives under diverse agent strategies while adhering the desired cost constraint. ASRL surpasses traditional learning-based and control-theoretic approaches, demonstrating adaptiveness and generalization across various conditions, such as number of agents and their parameters, thus enabling safe autonomy in dynamic multi-agent setting.

Why ASRL in multi-agent systems? CBFs are a valuable tool but their design and tuning is challenging with multiple agents. We enhance CBFs with adaptive coefficients, integrating them into a trainable architecture and optimize them to diverse behaviors and long-term objectives.

How does ASRL compare to existing approaches? ASRL retains benefits of CBF over learning methods, enabling efficient exploration while consistently adhering to cost limits. Compared to control-theoretic methods, our trainable model achieves superior performance and adaptability.

What are the limitations of ASRL? We focus on systems with a relative degree of 1. To consider higher-order systems and CBFs [42], it would be possible to introduce multiple coefficients in our approach. However, expanding the action space can make high-dimensional continuous control and exploration challenging. Also, ASRL does not assume full observability of other agents or explicit uncertainty quantification, thereby limiting its ability to guarantee safety in all scenarios. Finally, the assumption on CBF availability might restrict its applicability. Ongoing research explores alternative methods to scale to more complex dynamics and quantify uncertainty in pursuit of robust solutions.

VIII. Acknowledgements

L.B. was supported by DCRES. This project received partial support by NSF CCRI 1925587 and US DoT Safety21 UTC.

REFERENCES

[1] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2016.

[2] C. Dawson, S. Gao, and C. Fan. Safe control with learned certificates: A survey of neural lyapunov, barrier, and contraction methods for robotics and control. *IEEE Transactions on Robotics*, 2023.

[3] S. Yang, S. Chen, V. M. Preciado, and R. Mangharam. Differentiable safe controller design through control barrier functions. *IEEE Control Systems Letters*, 7:1207–1212, 2022.

[4] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3387–3395, 2019.

[5] Y. Emam, G. Notomista, P. Glotfelter, Z. Kira, and M. Egerstedt. Safe reinforcement learning using robust control barrier functions. *IEEE Robotics and Automation Letters*, (99):1–8, 2022.

[6] J. Choi, F. Castaneda, C. J. Tomlin, and K. Sreenath. Reinforcement learning for safety-critical control under model uncertainty, using control lyapunov functions and control barrier functions. *arXiv preprint arXiv:2004.07584*, 2020.

[7] J. Wang, S. Yang, Z. An, S. Han, Z. Zhang, R. Mangharam, M. Ma, and F. Miao. Multi-agent reinforcement learning guided by signal temporal logic specifications. *arXiv preprint arXiv:2306.06808*, 2023.

[8] X. Li and C. Belta. Temporal logic guided safe reinforcement learning using control barrier functions. *arXiv preprint arXiv:1903.09885*, 2019.

[9] W. Xiao, C. Belta, and C. G. Cassandras. Adaptive control barrier functions. *IEEE Transactions on Automatic Control*, 67(5):2267–2281, 2021.

[10] Z. Qin, K. Zhang, Y. Chen, J. Chen, and C. Fan. Learning safe multi-agent control with decentralized neural barrier certificates. *arXiv preprint arXiv:2101.05436*, 2021.

[11] W. Xiao, T.-H. Wang, R. Hasani, M. Chahine, A. Amini, X. Li, and D. Rus. Barriernet: Differentiable control barrier functions for learning of safe robot control. *IEEE Transactions on Robotics*, 2023.

[12] J. Zeng, B. Zhang, Z. Li, and K. Sreenath. Safety-critical control using optimal-decay control barrier function with guaranteed point-wise feasibility. In *2021 American Control Conference (ACC)*, pages 3856–3863. IEEE, 2021.

[13] X. Tan and D. V. Dimarogonas. Compatibility checking of multiple control barrier functions for input constrained systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 939–944. IEEE, 2022.

[14] M. Black and D. Panagou. Consolidated control barrier functions: Synthesis and online verification via adaptation under input constraints. *arXiv preprint arXiv:2304.01815*, 2023.

[15] R. Cheng, M. J. Khojasteh, A. D. Ames, and J. W. Burdick. Safe multi-agent interaction through robust control barrier functions with learned uncertainties. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 777–783. IEEE, 2020.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[17] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[18] A. Stooke, J. Achiam, and P. Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pages 9133–9143. PMLR, 2020.

[19] A. Ray, J. Achiam, and D. Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7(1):2, 2019.

[20] U. Borrmann, L. Wang, A. D. Ames, and M. Egerstedt. Control barrier certificates for safe swarm behavior. *IFAC-PapersOnLine*, 48(27):68–73, 2015.

[21] S. He, J. Zeng, B. Zhang, and K. Sreenath. Rule-based safety-critical control design using control barrier functions with application to autonomous lane change. In *2021 American Control Conference (ACC)*, pages 178–185. IEEE, 2021.

[22] M. O'Kelly, H. Zheng, D. Karthik, and R. Mangharam. F1tenth: An open-source evaluation environment for continuous control and reinforcement learning. In *NeurIPS 2019 Competition and Demonstration Track*, pages 77–89. PMLR, 2020.

[23] J. Ji, J. Zhou, B. Zhang, J. Dai, X. Pan, R. Sun, W. Huang, Y. Geng, M. Liu, and Y. Yang. Omnisafe: An infrastructure for accelerating safe reinforcement learning research. *arXiv preprint arXiv:2305.09304*, 2023.

[24] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[25] S. Fujimoto, H. Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[26] J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.

[27] Y. Liu, J. Ding, and X. Liu. Ipo: Interior-point policy optimization under constraints. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4940–4947, 2020.

[28] A. Sootla, A. I. Cowen-Rivers, T. Jafferjee, Z. Wang, D. H. Mguni, J. Wang, and H. Ammar. Sauté rl: Almost surely safe reinforcement learning using state augmentation. In *International Conference on Machine Learning*, pages 20423–20443. PMLR, 2022.

[29] M. M. Rahman and Y. Xue. Robust policy optimization in deep reinforcement learning. *arXiv preprint arXiv:2212.07536*, 2022.

[30] V. Sezer and M. Gokasan. A novel obstacle avoidance algorithm:"follow the gap method". *Robotics and Autonomous Systems*, 60(9):1123–1134, 2012.

[31] R Craig Coulter et al. *Implementation of the pure pursuit path tracking algorithm*. Carnegie Mellon University, The Robotics Institute, 1992.

[32] D. Ferguson, T. M. Howard, and M. Likhachev. Motion planning in urban environments. *Journal of Field Robotics*, 25(11-12):939–960, 2008.

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

[34] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause. Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30, 2017.

[35] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

[36] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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

[38] H. Parwana and D. Panagou. Recursive feasibility guided optimal parameter adaptation of differential convex optimization policies for safety-critical systems. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6807–6813. IEEE, 2022.

[39] H. Parwana, A. Mustafa, and D. Panagou. Trust-based rate-tunable control barrier functions for non-cooperative multi-agent systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 2222–2229. IEEE, 2022.

[40] L. Wang, A. D. Ames, and M. Egerstedt. Safety barrier certificates for collisions-free multirobot systems. *IEEE Transactions on Robotics*, 33(3):661–674, 2017.

[41] Y. Chen, A. Singletary, and A. D. Ames. Guaranteed obstacle avoidance for multi-robot operations with limited actuation: A control barrier function approach. *IEEE Control Systems Letters*, 5(1):127–132, 2020.

[42] W. Xiao and C. Belta. Control barrier functions for systems with high relative degree. In *2019 IEEE 58th conference on decision and control (CDC)*, pages 474–479. IEEE, 2019.