

Safe Reinforcement Learning using Data-Driven Predictive Control

Mahmoud Selim¹, Amr Alanwar², M. Watheq El-Kharashi¹, Hazem M. Abbas¹, Karl H. Johansson³

Abstract—Reinforcement learning (RL) algorithms can achieve state-of-the-art performance in decision-making and continuous control tasks. However, applying RL algorithms on safety-critical systems still needs to be well justified due to the exploration nature of many RL algorithms, especially when the model of the robot and the environment are unknown. To address this challenge, we propose a data-driven safety layer that acts as a filter for unsafe actions. The safety layer uses a data-driven predictive controller to enforce safety guarantees for RL policies during training and after deployment. The RL agent proposes an action that is verified by computing the data-driven reachability analysis. If there is an intersection between the reachable set of the robot using the proposed action, we call the data-driven predictive controller to find the closest safe action to the proposed unsafe action. The safety layer penalizes the RL agent if the proposed action is unsafe and replaces it with the closest safe one. In the simulation, we show that our method outperforms state-of-the-art safe RL methods on the robotics navigation problem for a Turtlebot 3 in Gazebo and a quadrotor in Unreal Engine 4 (UE4).

Index Terms—Reinforcement learning, robot safety, task and motion planning.

I. INTRODUCTION

Safety of Cyber-Physical Systems (CPSs) has always been of great interest to researchers [1]. This imposes a great challenge for reinforcement learning controllers due to the exploration nature of many reinforcement learning algorithms [2]. This is especially true when the environment and/or system models are both time-varying or subject to noise [3]. Recently, many techniques have been proposed to mitigate the safety problems of RL algorithms. An RL agent is considered to be safe if it meets an ergodicity requirement that it can reach each state it visits from any other state it visits, allowing for reversible errors [4], [5]. To justify the widespread deployment of RL controllers, we need to ensure the safety of RL algorithms during training and after deployment [6]. In this work, we propose a safety layer to address the safety challenges of RL algorithms in a totally data-driven way when the robot model is unknown. We utilize data-driven predictive control [7] to impose strict safety constraints on RL agents. This is considered a step towards making data-driven methods more convenient and practical for safety-critical systems.

A. Related Work

Safe RL aims to maximize an objective function while respecting safety constraints. Many methods have been previously proposed to address the problem of safety for RL agents. One popular approach imposes constraints on the expected return [8], [9]. Other include risk measures [10], [11], [12], [13] or impose constraints on the Markov Decision

Process (MDP) [14] or avoid regions in the MDP where the safety constraints may be violated [15], [16]. We note that these approaches are different from just penalizing the agent with large negative rewards. This type of penalty can lead to undesired behavior from the agent and may cause the training to be unstable.

Another major approach to safe RL modifies the exploration process to filter out unsafe actions taken by the RL agent. This can be achieved by two methods. The first one uses a risk-directed exploration, and the second depends on the use of external knowledge [5]. Risk-directed exploration usually uses a risk-measure to determine the probability of selecting different actions during the exploration process [5], [17]. External knowledge, on the other hand, can be done by providing a set of demonstrations [18], or the use of control theoretic approaches. Control theoretic approaches usually require a partial or a complete system model to predict whether the system will reach an unsafe state and prevent it. For example, one can utilize control barrier functions and control Lyapunov functions to guarantee the system safety or stability [19], [20], [21], [22]. Others use reachability-based methods to calculate the reachable sets of the system. Safety is then achieved when the reachable set of the system lies within the safe region. However, if they don't, action correction techniques can be used, such as sampling to find a safe action or projecting the RL action into the safe set. This can be done efficiently in case of parameterized reachability as in [23].

One final remark, our proposed method is closely related to [24], [25]. They both use data collected offline to enforce safety constraints on the system. However, the former uses a first-order approximation shown to be quite conservative. The latter gives no regard to how much the adjusted action is close to the RL agents' action of choice. The proposed approach, however, tries to be as little invasive as possible to the RL agents' action of choice, similar to [26].

Trajectory or path planning is considered a fundamental problem in robotics. It aims to find the shortest and most obstacle-free path from the start to the goal state. The path can be a set of states (position and orientation) or waypoints. Many methods have been proposed to address the planning problems in robotics. They are mainly classified into two categories: classical approaches and learning-based approaches. Classical approaches can be further divided into grid-based search algorithms and sampling-based search algorithms. Grid-based search algorithms usually find a path using a minimum travel cost like Dijkstra's algorithm and A*. Grid-based algorithms are guaranteed to find the shortest path given enough computation time and resources. However, in most real-life scenarios, finding a reasonable path or a trajectory in an acceptable computation time is preferable. Sampling-based algorithms

¹Ain Shams University, Cairo, Egypt. ²Jacobs University, Bremen, Germany. ³KTH Royal Institute of Technology, Stockholm, Sweden. Corresponding author: mahmoud.selim@eng.asu.edu.eg.

take this approach. That is, they sample the configuration space to reach a reasonable solution. Sampling algorithms such as RRT* [27, Ch. 5] are probabilistically complete and are suitable for both low and high-dimensional search spaces. Recent approaches to path planning rely heavily on learning-based approaches. Many Deep Reinforcement Learning (DRL) techniques have been developed to address the planning problem, especially in uncertain environments. Some of these approaches depend on model-based reinforcement learning for planning [28]. Others utilize hierarchical reinforcement learning for planning tasks [29], [30]. A clear advantage of these methods is that, given enough training samples, they can be optimized while remaining computationally efficient.

Limitations. We assume a discrete-time setting and model the system as a linear model. We leave the continuous-time setting in reachability analysis [31]), non-linear system models, and the amount of data required to guarantee safety for future work.

Contributions. The main contributions of this paper can be summarized as follows:

- 1) We propose a safety filter using data-driven predictive control to enforce safety constraints while being as little invasive as possible to the RL agents' choice of action when the model of the robot is unknown.
- 2) We demonstrate the proposed method for the motion planning problem. We show through experiments that the proposed method is, indeed, able to guarantee safety. We also compare the proposed method against a baseline RL agent, Safe Exploration in Continuous Action Spaces (SECAS) [24], and Safe Advantage-based Intervention for Learning policies with Reinforcement (SAILR) [32].

Next, in Section II, we provide the problem statement and formulate the safe RL problem. Section III discusses the proposed method, and Section IV goes through the evaluation and discusses the result. Finally, Section V presents concluding remarks and discusses future work.

II. PRELIMINARIES AND PROBLEM FORMULATION

We now define the notation used throughout the paper, set representations for our reachability analysis, the system and its reachable sets, and our assumptions about data and noise.

A. Notation and Set Representations

Natural numbers are denoted by \mathbb{N} , and n -dimensional real numbers by \mathbb{R}^n . The element at row i and column j of a matrix A is denoted by $(A)_{i,j}$, the column j of A by $(A)_{:,j}$. The pseudoinverse of a matrix A is denoted by A^\dagger . For simplicity, the Minkowski sum, defined by $\mathcal{Z}_1 \oplus \mathcal{Z}_2 = \{z_1 + z_2 \mid z_1 \in \mathcal{Z}_1, z_2 \in \mathcal{Z}_2\}$, is denoted by $+$ instead of \oplus as the type can be determined from the context. Similarly, we use $\mathcal{Z}_1 - \mathcal{Z}_2$ to denote $\mathcal{Z}_1 + -1\mathcal{Z}_2$. The Cartesian product is given by $A \times B = \{[a^\top \ b^\top]^\top \mid a \in A, b \in B\}$.

Reachable sets are represented using zonotopes.

Definition 1. (Zonotope [33]) Given a center $c_Z \in \mathbb{R}^n$ and $\gamma_Z \in \mathbb{N}$ generator vectors in a generator matrix $G_Z = [g_Z^{(1)}, \dots, g_Z^{(\gamma_Z)}] \in \mathbb{R}^{n \times \gamma_Z}$, a zonotope is the set

$$\mathcal{Z} = \left\{ x \in \mathbb{R}^n \mid x = c_Z + \sum_{i=1}^{\gamma_Z} \beta^{(i)} g_Z^{(i)}, -1 \leq \beta^{(i)} \leq 1 \right\},$$

where $\beta^{(i)}$ is called the zonotope's factor. We use the shorthand notation $\mathcal{Z} = \mathcal{Z}(c_Z, G_Z)$ for a zonotope.

Given a linear map L , one can show that $L\mathcal{Z} = \mathcal{Z}(Lc_Z, LG_Z)$. Given two zonotopes $\mathcal{Z}_1 = \langle c_{Z_1}, G_{Z_1} \rangle$ and $\mathcal{Z}_2 = \langle c_{Z_2}, G_{Z_2} \rangle$, their Minkowski sum is computed by

$$\mathcal{Z}_1 + \mathcal{Z}_2 = \left\langle c_{Z_1} + c_{Z_2}, [G_{Z_1}, G_{Z_2}] \right\rangle. \quad (1)$$

We compute the Cartesian product of two zonotopes \mathcal{Z}_1 and \mathcal{Z}_2 by

$$\mathcal{Z}_1 \times \mathcal{Z}_2 = \left\langle \begin{bmatrix} c_{Z_1} \\ c_{Z_2} \end{bmatrix}, \begin{bmatrix} G_{Z_1} & 0 \\ 0 & G_{Z_2} \end{bmatrix} \right\rangle. \quad (2)$$

We also make use of matrix zonotopes to represent families of matrices:

Definition 2. (Matrix zonotope [34, p. 52]) Given a center matrix $C_M \in \mathbb{R}^{n \times T}$ and $\gamma_M \in \mathbb{N}$ generator matrices $\tilde{G}_M = [G_M^{(1)}, \dots, G_M^{(\gamma_M)}] \in \mathbb{R}^{n \times (T\gamma_M)}$, a matrix zonotope is defined as

$$\mathcal{M} = \left\{ X \in \mathbb{R}^{n \times T} \mid X = C_M + \sum_{i=1}^{\gamma_M} \beta^{(i)} G_M^{(i)}, -1 \leq \beta^{(i)} \leq 1 \right\}. \quad (3)$$

Definition 3. (Interval Matrix [34, p.42]) An interval matrix \mathcal{I} specifies the interval of all possible values for each matrix element between the left limit \underline{I} and right limit \bar{I} :

$$\mathcal{I} = [\underline{I}, \bar{I}], \quad \underline{I}, \bar{I} \in \mathbb{R}^{r \times c} \quad (4)$$

B. System Dynamics and Reachable Sets

We assume a discrete time linear system model:

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + w(t), \\ y(t) &= Cx(t) + v(t). \end{aligned} \quad (5)$$

with the system matrices $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$, state $x(t) \in \mathbb{R}^n$, input $u(t) \in \mathbb{R}^m$, process noise $w(t) \in \mathbb{R}^n$, and measurement noise $v(t) \in \mathbb{R}^p$. We assume that the states of the system are measurable and compact, i.e., the system output matrix is given by $C = I$, and thus the measured output is $y(t) \in \mathbb{R}^n$. Given that $C = I$, we call the $y(t)$ the state which is corrupted by extra noise $v(t)$. The input and output constraints are given by

$$\begin{aligned} u(t) &\in \mathcal{U}_t \subset \mathbb{R}^m, \\ y(t) &\in \mathcal{Y}_t \subset \mathbb{R}^n. \end{aligned} \quad (6)$$

Reachability analysis computes the reachable set of a system, which is the set of states $y(t)$ which can be reached given a set of uncertain initial states $\mathcal{Y}_0 \subset \mathbb{R}^n$ and a set of possible inputs $\mathcal{Z}_{u,t}$ at each time step $t \in \mathbb{N}$. More formally, we define the reachable set at time $N \in \mathbb{N}$ as follows:

Definition 4. The reachable set R_t after N time steps, subject to a sequence of inputs $u(t) \in \mathcal{Z}_{u,t}$, noise $w(t) \in \mathcal{Z}_w$, measurement noise $v(t) \in \mathcal{Z}_v$, $\forall t \in \{0, \dots, N-1\}$, and initial set $Y_0 \in \mathbb{R}^n$, is the set

$$\begin{aligned} R_N &= \{y(N) \in \mathbb{R}^n \mid x(t+1) = Ax(t) + Bu(t) + w(t), \\ &\quad y(t) = x(t) + v(t), x_0 \in Y_0, u(t) \in \mathcal{Z}_{u,t}, v(t) \in \mathcal{Z}_v, \\ &\quad \text{and } w(t) \in \mathcal{Z}_w, \forall t = 0, \dots, N-1\}. \end{aligned} \quad (7)$$

In this work, we assume unknown system dynamics. However, we still seek to over-approximate the reachable sets of the system using data-driven reachability [35]. This way, we show that we are able to enforce safety constraints on a black-box system leveraging only data-driven methods.

To enable safety guarantees, we leverage the notion of failsafe maneuvers from mobile robotics [36]. Note that many real robots have a braking safety controller available, similar to the notion of an invariant set [37], [31]. Also, failsafe maneuvers exist even when a robot cannot remain stationary, such as loiter circles for aircraft [38].

We do not assume a distribution for the noise. However, we require that the noise obeys the following assumptions for numerical tractability and robustness guarantees.

Assumption 1. We assume that $w(t)$ is bounded by noise zonotope $\mathcal{Z}_w = \mathcal{Z}(\mathbf{c}_w, \mathbf{G}_w)$.

Assumption 2. We assume that $v(t)$ is bounded by a zonotope $\mathcal{Z}_v = \mathcal{Z}(\mathbf{c}_v, \mathbf{G}_v)$. Furthermore, we assume that the one-step propagation $Av(t)$ is bounded by a zonotope $Av(t) \in \mathcal{Z}_{Av}$ for all time steps.

In this work, unsafe regions of state space, or *obstacles* are denoted by as $Y_{\text{obs}} \subset \mathbb{R}^n$, and since the focus of this work is not motion forecasting of other actors in the environment, obstacles are assumed to be static and not to move (yet they change from one episode to another). We emphasize, however, that this work can be extended to handle dynamic environments by using frameworks that use the reachability of other agents [39], [40].

C. Safe RL Problem Formulation

We formulate the safe RL problem as follows. Let $Y_{\text{goal}} \subset \mathbb{R}^n$ denote a goal the agent aims to reach while avoiding the obstacles (unsafe sets) $Y_{\text{obs}} \subset \mathbb{R}^n$. Let $n_{\text{total}} \in \mathbb{N}$ denote a total number of time steps in the training procedure, and n_{iter} denote the number of steps per episode. At time t , let $\hat{y}(t) \in \mathbb{R}^{n_{\text{RL}}}$ denote the state of the RL agent (which contains the state $y(t)$ of the robot, plus additional information such as sensor measurements and previous actions). Let $u(t)$ denote the action chosen by the RL agent at time t . Let

$$r : (\hat{y}(t), Y_{\text{goal}}, u(t)) \mapsto \mathbb{R} \quad (8)$$

denote a real-valued reward function. We seek to learn a policy

$$\pi^* : \hat{y}(t) \mapsto u(t) \quad (9)$$

which maximizes the cumulative reward. Furthermore, we require that, if we roll out the trajectory given the policy π from any initial condition in the set Y_0 , then the trajectory is safe.

III. SAFE REINFORCEMENT LEARNING

We propose a new framework to enforce safety constraints on the RL agents without previous knowledge of the environment or the robot model apriori. The idea is to introduce a safety layer that filters out any unsafe action choice taken by the RL agent in a totally data-driven way while being as little invasive as possible for the agent's choice of action. Since the framework uses data-driven methods to enforce safety, the proposed method can handle time-varying systems by updating

Algorithm 1: Safe Reinforcement Learning

```

1 Initialize the RL agent with a random policy  $\pi_\theta$ , empty
  replay buffer  $B$ , number of steps per episode  $n_{\text{iter}}$ ,
  and a safe plan  $\mathbf{p}_0$ 
2 for each episode do
3   initialize RL task.
4    $\hat{y}(1) \leftarrow$  observe initial environment state.
5   for  $k = 1 : n_{\text{iter}}$  do
6      $u(k) \leftarrow$  sample RL policy
7      $\hat{R}_k \leftarrow \mathcal{Z}(y(k), \mathbf{0})$  // init. reachable set
8      $\mathbf{p}_k \leftarrow$  enforce_safety( $\hat{R}_k, u(k), y(k)$ ) // Alg. 2
9     if  $\mathbf{p}_k == \emptyset$  then
10      // if there is no found safe plan
11      execute failsafe maneuver; continue
12      $u(k) \leftarrow$  get first (safe) action from  $\mathbf{p}_k$ 
13      $r_k \leftarrow \rho(\hat{y}(k), u(k))$  // get reward
14      $\hat{y}(k+1) \leftarrow$  observe new state
15     add ( $\hat{y}(k), u(k), r_k, \hat{y}(k+1)$ ) to  $B$ 
16     train the RL agent  $\pi_\theta$  using a batch from  $B$ 

```

the data matrices to accommodate changes to the environment or the system model.

The proposed safety filter is summarized in Algorithm 1. It uses a receding-horizon strategy to create a new safe plan \mathbf{p}_k in each k receding-horizon motion planning iteration. Consider a single planning iteration k (Lines 4–16). Suppose the robot has previously created a safe plan \mathbf{p}_{k-1} (such as staying stopped indefinitely). At the beginning of each iteration, our framework samples the RL policy for a new action $u(k)$ (Line 6). Next, it enforces safety on the action chosen by the RL agent by solving a data-driven optimal control problem as in Algorithm 2. If Algorithm 2 fails to find a safe plan, it executes a failsafe maneuver instead (Lines 9–11). Then, we apply the first action of the safe plan to the environment, get a reward, observe the environment, and train the RL agent from the replay buffer (Lines 12–16). We note that the proposed algorithm can be used during both training and deployment. That is, the safety layer can operate even for an untrained policy while initializing π_θ with random weights.

Data-driven predictive control solves an optimization problem to find a reachable set (represented by a zonotope) under some constraints [7]. It utilizes data-driven reachability analysis [35], [41] under the hood to overapproximate the reachable sets of an unknown system from noisy data collected offline. In this work, we assume a linear system model.

We consider q input-state trajectories of lengths $T_i \in \mathbb{N}$, $i = 1, \dots, q$, with total duration $T_{\text{total}} = \sum_i^q T_i$. We denote the data as $(y_k^{(i)})_{k=0}^{T_i}$, $(u_k^{(i)})_{k=0}^{T_i-1}$, $i = 1, \dots, q$. To further ease the, we rearrange the data in the following matrices:

$$Y_- = \begin{bmatrix} y_0^{(1)}, \dots, y_{T_1-1}^{(1)}, \dots, y_0^{(q)}, \dots, y_{T_q-1}^{(q)} \end{bmatrix}, \quad (10a)$$

$$Y_+ = \begin{bmatrix} y_1^{(1)}, \dots, y_{T_1}^{(1)}, \dots, y_1^{(q)}, \dots, y_{T_q}^{(q)} \end{bmatrix}, \quad (10b)$$

$$U_- = \begin{bmatrix} u_0^{(1)}, \dots, u_{T_1-1}^{(1)}, \dots, u_0^{(q)}, \dots, u_{T_q-1}^{(q)} \end{bmatrix}, \quad (10c)$$

and similarly for the unknown noise realizations V_- , V_+ and W_- . It follows directly from Assumptions 1 and 2 that the noise realizations are within corresponding matrix noise zonotopes V_- , $V_+ \in \mathcal{M}_v$, $W_- \in \mathcal{M}_w$ and $AV_- \in \mathcal{M}_{Av}$ as shown in [41].

Next, we compute a set of models consistent with the collected data which is utilized to compute the data-driven reachable sets in the following lemmas.

Lemma 1 ([7]). *Given input-output trajectories $D = \{U_-, Y\}$ of the system (5), then*

$$\mathcal{M}_\Sigma = (Y_+ - \mathcal{M}_w - \mathcal{M}_v + \mathcal{M}_{Av}) \begin{bmatrix} Y_- \\ U_- \end{bmatrix}^\dagger \quad (11)$$

contains all matrices $\begin{bmatrix} A & B \end{bmatrix}$ that are consistent with the data and noise bounds.

Lemma 2 ([7]). *Given input-output trajectories $D = \{U_-, Y\}$ of the system in (5), then*

$$\hat{\mathcal{R}}_{t+1} = \mathcal{M}_\Sigma(\hat{\mathcal{R}}_t \times \mathcal{Z}_{u,t}) + \mathcal{Z}_w + \mathcal{Z}_v - \mathcal{Z}_{Av}, \quad (12)$$

contains the reachable set, i.e., $\hat{\mathcal{R}}_{t+1} \supseteq \mathcal{R}_{t+1}$ where $\hat{\mathcal{R}}_0 = \mathcal{Z}(y(0), 0)$, and $\mathcal{Z}_{u,t} = \mathcal{Z}(u(t), 0)$.

To enforce safety, we must generate a plan instead of a single action. The reason behind this is that safety can't be enforced for a single action as the actuation of the system is finite (e.g., a robot needs N time steps to brake). Algorithm 2 enforces safety while being as little invasive as possible to the RL agents' choice of action. This is done by solving a data-driven optimal control problem as in Line 2 of Algorithm 2. This optimization problem is solved under some constraints (avoiding obstacles). First, we apply data-driven reachability on the state zonotope $\hat{\mathcal{R}}_{t+k+1|t}$ in (13b). The computed reachable set $\hat{\mathcal{R}}_{t+k+1|t}$, which includes the output in line (13e), is ensured to satisfy the output constraints \mathcal{Y}_{t+k+1} in order to guarantee the safety in (13d). The initial state in the optimization problem is set to be the same as the robot state in (13f).

Remark 1. *The constraints (13d) is implemented by converting both zonotopes into intervals and ensuring that*

$$\hat{\mathcal{R}}_{u,t+k+1|t} \leq \mathcal{Y}_{u,t+k+1}, \quad \hat{\mathcal{R}}_{l,t+k+1|t} \geq \mathcal{Y}_{l,t+k+1}$$

where $\hat{\mathcal{R}}_{u,t+k+1|t}$ and $\hat{\mathcal{R}}_{l,t+k+1|t}$ are the upper and lower bounds of $\hat{\mathcal{R}}_{t+k+1|t}$, and $\mathcal{Y}_{u,t+k+1}$ and $\mathcal{Y}_{l,t+k+1}$ are the upper and lower bounds of \mathcal{Y}_{t+k+1} .

We conclude this section by formalizing the safety guarantees of our proposed algorithm.

Theorem 1. *Suppose the assumptions on the robot and environment from Section II all hold, and, at time $k = 0$, the robot is at safe state. Suppose also that, at each time $k > 0$, the robot rolls out a new \mathbf{p}_k , and adjusts the plan using Algorithm 2. Then, the robot is guaranteed to be safe at all times $k \geq 0$.*

Proof. The proof can be done by induction. At the time 0, the agent can apply a braking action u_{brk} to stay safe for all time. Assume a safe plan exists at time $k \in \mathbb{N}$. Then, at any timestep $k + 1$, in case of failure of Algorithm 2 to find a safe plan, and since the plan is always larger than the number

Algorithm 2: Enforce Safety Using Data Driven Predictive Control

Input: action $u(k)$, state $y(k)$, obstacles Y_{obs} , initial reachable set $\hat{\mathcal{R}}_k$, time limit t_{max} , planning horizon n_{plan} , and process noise zonotope \mathcal{Z}_w

Output: Safe plan \mathbf{p}_k , if available

- 1 Set the reference input action $r \leftarrow u(k)$, and the obstacles free area to \mathcal{Y}_{i+k+1}
 - 2 Solve the optimization problem (for $i = 0, \dots, n_{\text{plan}}$) where the input action zonotope $\mathcal{Z}_{u,k} \leftarrow \mathcal{Z}(u_{i+k|k}, 0)$

$$\min_{u,y} \|u_{k|k} - r\|^2 \quad (13a)$$

s.t. $\hat{\mathcal{R}}_{i+k+1|k} = \mathcal{M}_\Sigma(\hat{\mathcal{R}}_{i+k|k} \times \mathcal{Z}_{u,k}) + \mathcal{Z}_w + \mathcal{Z}_v - \mathcal{Z}_{Av}, \quad (13b)$

$$u_{i+k|k} \in \mathcal{U}_{i+k}, \quad (13c)$$

$$\hat{\mathcal{R}}_{i+k+1|k} \subseteq \mathcal{Y}_{i+k+1}, \quad (13d)$$

$$y_{i+k+1|k} \in \hat{\mathcal{R}}_{i+k+1|k}, \quad (13e)$$

$$y_{k|k} = y(k) \quad (13f)$$
 - 3 if all $\hat{\mathcal{R}}_j \cap Y_{\text{obs}} = \emptyset$ then
 - 4 | **return** $\mathbf{p}_k = (u_j)_{j=k}^{n_{\text{plan}}}$ // found a safe plan
 - 5 else
 - 6 | **return** $\mathbf{p}_k = \emptyset$ // No safe plan was found
-

of steps required by the robot to brake to a stop fully, the robot always has a failsafe braking maneuver that enables it to stop indefinitely; otherwise, if a new plan is found, the plan is safe for two reasons. First, the reachability algorithm is guaranteed to contain the true reachable set of the system [35, Theorem 2] because the noise zonotopes bound noise as in Assumptions 1 and 2. Second, when adjusting an unsafe action with Algorithm 2, the algorithm will only return a safe plan in which the first action is as close as possible to the RL action of choice while enforcing the hard constraints of safety. We emphasize that in order for a plan to be deemed safe, the plan length must be larger than the number of steps needed for the robot to brake to a full stop. \square

IV. EXPERIMENTS

We evaluate the proposed frameworks on two motion planning tasks in simulation: A Turtlebot in Gazebo and a Quadrotor in Unreal Engine 4 (UE4). Fig. 1 shows our simulation environments. The simulation shows that the proposed framework shields the RL agents from unsafe actions while maximizing the objective. Next, we describe our simulation environments in detail.

A. Simulation Environments Setup

1) Turtlebot Setup

We use Turtlebot 3 in the Gazebo simulator. The robot has a longitudinal velocity in $[0, 0.25]$ m/s, and an angular velocity in $[-0.5, 0.5]$ rad/s. We clip the upper longitudinal velocity at 0.25 m/s to be consistent with the real robot dynamics. The robot is also provided with wheel encoders, an IMU to estimate the robot's speed and position, and a planar lidar with 18 laser measurements is evenly spanning 180° . In each episode, the robot is spawned randomly in a safe spot. It's required to reach

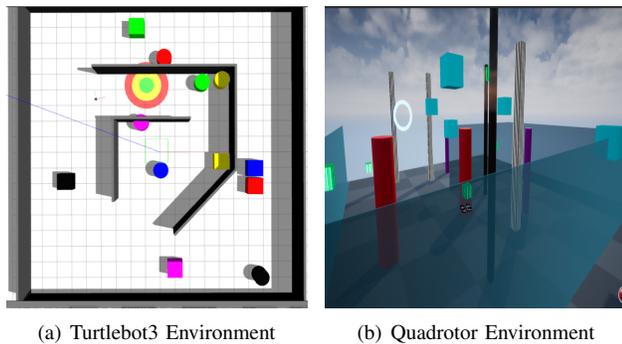


Fig. 1: Evaluation Environments.

the goal on the map denoted by the green circle in Fig. 1(a). The goal is chosen at a random location in each episode, and the obstacles are spawned randomly in each episode as well.

2) Quadrotor

The quadrotor is simulated in Unreal Engine 4 (UE4). They have a similar dynamics with a velocity ranging from $[-5, -5, -5]$ m/s to $[5, 5, 5]$ m/s. Both are an IMU to estimate the rotors' position and velocity and are provided with a multi-channel 3D lidar. The lidar spans a range of 360° horizontally and a range of 60° vertically. Again, the environments' obstacles and goals are spawned randomly at the beginning of each episode. The goal of the quadrotor is to reach a goal position while avoiding the obstacles.

B. RL Agent Setup

1) Turtlebot Environment

We use TD3 [42] as our RL agent. The state vector $\hat{y} \in \mathbb{R}^{26}$ consists of the state of the physical robot $\in \mathbb{R}^8$ (position and orientation) estimated from the noisy wheel encoders and IMU, and the lidar measurements $\in \mathbb{R}^{18}$.

2) Quadrotor Setup

Similar to the Turtlebot example, we use TD3 as our RL agent. The state vector $\hat{y} \in \mathbb{R}^{189}$ consists of the state of the physical robot $\in \mathbb{R}^9$ (position only) estimated from the noisy IMU, and the lidar measurements $\in \mathbb{R}^{180}$.

C. Results and Discussion

The results are summarized in Table I, and Fig. 2. We see that the proposed method could achieve the highest reward and outperform all the other methods. As for the safety constraints, we see from Table I that both the proposed method and SECAS could impose hard safety constraints (i.e., having 0 collision rate). However, both SAILR and TD3 agents couldn't have full constraint satisfaction. Regarding the goal rate and the robot speed, our proposed method outperforms all the other methods without being overly conservative. Finally, regarding computation time, we see that the vanilla TD3 RL agent outperformed all the other methods due to the absence of any safety modules in the framework. However, we note that the proposed method is real-time even when dealing with large data matrices since the calculations can be done efficiently on GPUs.

V. CONCLUSION

This paper proposes a framework for safe reinforcement learning. The proposed framework leverages data-driven methods to provide safety guarantees. The framework was evaluated

on a path planning task for Turtlebot 3 and a quadrotor, where we proved mathematically and experimentally that the proposed method could indeed provide safety under reasonable assumptions. For future work, we will explore continuous-time settings, reducing the conservativeness of our reachability analysis, and minimizing the amount of data needed to guarantee safety.

ACKNOWLEDGEMENT

This paper has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 830927.

REFERENCES

- [1] A. Banerjee, K. K. Venkatasubramanian, T. Mukherjee, and S. K. S. Gupta, "Ensuring safety, security, and sustainability of mission-critical cyber-physical systems," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 283–299, 2012.
- [2] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 135.
- [3] O. Mihatsch and R. Neuneier, "Risk-sensitive reinforcement learning," *Machine learning*, vol. 49, no. 2, pp. 267–290, 2002.
- [4] T. M. Moldovan and P. Abbeel, "Safe exploration in markov decision processes," 2012.
- [5] J. Garcia and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.
- [6] K. Asadi, D. Misra, S. Kim, and M. L. Littman, "Combating the compounding-error problem with a multi-step model," 2019.
- [7] A. Alanwar, Y. Stürz, and K. H. Johansson, "Robust data-driven predictive control using reachability analysis," *European Journal of Control*, p. 100666, 2022.
- [8] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," 2017.
- [9] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," in *International Conference on Learning Representations*, 2018.
- [10] A. Tamar, Y. Glassner, and S. Mannor, "Policy gradients beyond expectations: Conditional value-at-risk," 04 2014.
- [11] M. Heger, "Consideration of risk in reinforcement learning," 03 1994.
- [12] P. Geibel and F. Wysotzki, "Risk-sensitive reinforcement learning applied to control under constraints," *Journal of Artificial Intelligence Research*, vol. 24, pp. 81–108, jul 2005.
- [13] Y. Shen, M. J. Tobia, T. Sommer, and K. Obermayer, "Risk-sensitive reinforcement learning," *Neural Computation*, vol. 26, no. 7, pp. 1298–1328, 2014.
- [14] H. B. Ammar, R. Tutunov, and E. Eaton, "Safe policy search for lifelong reinforcement learning with sublinear regret," 2015.
- [15] M. Turchetta, F. Berkenkamp, and A. Krause, "Safe exploration in finite markov decision processes with gaussian processes," 2016.
- [16] B. Eysenbach, S. Gu, J. Ibarz, and S. Levine, "Leave no trace: Learning to reset for safe and autonomous reinforcement learning," 2017.
- [17] C. Gehring and D. Precup, "Smart exploration in reinforcement learning using absolute temporal difference errors," in *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems*, 2013, p. 1037–1044.
- [18] P. Abbeel and A. Y. Ng, "Exploration and apprenticeship learning in reinforcement learning," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 1–8.
- [19] T. G. Molnar, R. K. Cosner, A. W. Singletary, W. Ubellacker, and A. D. Ames, "Model-free safety-critical control for robotic systems," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 944–951, 2021.

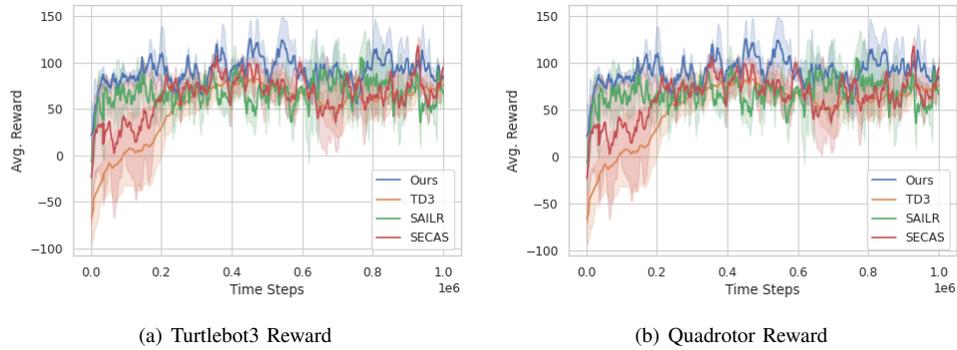


Fig. 2: Average reward over time of Our method (blue), SAILR [32], SECAS [24], and a vanilla TD3 baseline for each of our experiments.

TABLE I: Numerical experiment results (best values in bold)

	Turtlebot				Quadrotor			
	ours	SAILR	SECAS	Baseline	ours	SAILR	SECAS	Baseline
Goal Rate [%]	55	50	53	38	68	60	61	46
Collision Rate [%]	0.0	4.9	0.0	17	0.0	5.6	0.0	22
Mean/Max Speed [m/s]	.06 / 0.25	.04 / 0.17	.06 / 0.21	.08 / .25	3.5 / 8.6	3.2 / 8.2	3.2 / 8.3	3.7 / 8.6
Mean Reward	84	69	66	58	86	78	72	63
Mean \pm Std. Dev.	80.41 \pm 15.0	29.8 \pm 11.8	21.4 \pm 11.8	8.0 \pm 7.19	80.9 \pm 17.27	42.22 \pm 17.58	32.7 \pm 28.05	12.7 \pm 7.5
Computation Time [ms]								

- [20] P. Liu, D. Tateo, H. B. Ammar, and J. Peters, “Robot reinforcement learning on the constraint manifold,” in *Conference on Robot Learning*. PMLR, 2022, pp. 1357–1366.
- [21] M. Jankovic, “Robust control barrier functions for constrained stabilization of nonlinear systems,” vol. 96, pp. 359–367, 10 2018.
- [22] P. Seiler, M. Jankovic, and E. Hellstrom, “Control barrier functions with unmodeled dynamics using integral quadratic constraints,” 2021. [Online]. Available: <https://arxiv.org/abs/2108.10491>
- [23] N. Kochdumper, H. Krasowski, X. Wang, S. Bak, and M. Althoff, “Provably safe reinforcement learning via action projection using reachability analysis and polynomial zonotopes,” *arXiv preprint arXiv:2210.10691*, 2022.
- [24] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa, “Safe exploration in continuous action spaces.”
- [25] M. Selim, A. Alanwar, S. Kousik, G. Gao, M. Pavone, and K. H. Johansson, “Safe reinforcement learning using black-box reachability analysis,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 665–10 672, 2022.
- [26] K. P. Wabersich and M. N. Zeilinger, “A predictive safety filter for learning-based control of constrained nonlinear dynamical systems,” *Automatica*, vol. 129, p. 109597, 2021.
- [27] S. M. LaValle, *Planning algorithms*. Cambridge university press, 2006.
- [28] S. Guo, X. Zhang, Y. Zheng, and Y. Du, “An autonomous path planning model for unmanned ships based on deep reinforcement learning,” *Sensors*, vol. 20, no. 2, 2020.
- [29] J. B. Hamrick, A. L. Friesen, F. Behbahani, A. Guez, F. Viola, S. Witherspoon, T. Anthony, L. Buesing, P. Veličković, and T. Weber, “On the role of planning in model-based deep reinforcement learning,” 2020. [Online]. Available: <https://arxiv.org/abs/2011.04021>
- [30] J. Li, C. Tang, M. Tomizuka, and W. Zhan, “Hierarchical planning through goal-conditioned offline reinforcement learning,” 2022.
- [31] Y. S. Shao, C. Chen, S. Kousik, and R. Vasudevan, “Reachability-based trajectory safeguard (RTS): A safe and fast reinforcement learning safety layer for continuous control,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3663–3670, 2021.
- [32] N. Wagener, B. Boots, and C.-A. Cheng, “Safe reinforcement learning using advantage-based intervention,” *arXiv preprint arXiv:2106.09110*, 2021.
- [33] W. Kühn, “Rigorously computed orbits of dynamical systems without the wrapping effect,” *Computing*, vol. 61, no. 1, pp. 47–67, 1998.
- [34] M. Althoff, “Reachability analysis and its application to the safety assessment of autonomous cars,” Ph.D. dissertation, Technische Universität München, 2010.
- [35] A. Alanwar, A. Koch, F. Allgöwer, and K. H. Johansson, “Data-driven reachability analysis using matrix zonotopes,” in *Learning for Dynamics and Control*. PMLR, 2021, pp. 163–175.
- [36] S. Kousik, S. Vaskov, F. Bu, M. Johnson-Roberson, and R. Vasudevan, “Bridging the gap between safety and real-time performance in receding-horizon trajectory design for mobile robots,” *The International Journal of Robotics Research*, vol. 39, no. 12, pp. 1419–1469, 2020.
- [37] T. Lew, A. Sharma, J. Harrison, A. Bylard, and M. Pavone, “Safe active dynamics learning and control: A sequential exploration-exploitation framework,” *IEEE Transactions on Robotics*, 2022, in Press.
- [38] D. Fridovich-Keil, J. F. Fisac, and C. J. Tomlin, “Safely probabilistically complete real-time planning and exploration in unknown environments,” in *IEEE International Conference on Robotics and Automation*, 2019, pp. 7470–7476.
- [39] K. Leung, E. Schmerling, M. Zhang, M. Chen, J. Talbot, J. C. Gerdes, and M. Pavone, “On infusing reachability-based safety assurance within planning frameworks for human–robot vehicle interactions,” *The International Journal of Robotics Research*, vol. 39, no. 10-11, pp. 1326–1345, 2020.
- [40] S. Vaskov, H. Larson, S. Kousik, M. Johnson-Roberson, and R. Vasudevan, “Not-at-fault driving in traffic: A reachability-based approach,” in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 2785–2790.
- [41] A. Alanwar, A. Koch, F. Allgöwer, and K. H. Johansson, “Data-driven reachability analysis from noisy data,” *arXiv preprint arXiv:2105.07229*, 2021.
- [42] S. Dankwa and W. Zheng, “Twin-delayed ddpq: A deep reinforcement learning technique to model a continuous movement of an intelligent robot agent,” in *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*, 2019, pp. 1–5.