Verification of Neural Reachable Tubes via Scenario Optimization and Conformal Prediction

Albert Lin ALBERT.K.LIN@USC.EDU and **Somil Bansal** SOMILBAN@USC.EDU
Department of Electrical and Computer Engineering, University of Southern California, CA, USA

Editors: A. Abate, K. Margellos, A. Papachristodoulou

Abstract

Learning-based approaches for controlling safety-critical autonomous systems are rapidly growing in popularity; thus, it is important to provide rigorous and robust assurances on their performance and safety. Hamilton-Jacobi (HJ) reachability analysis is a popular formal verification tool for providing such guarantees, since it can handle general nonlinear system dynamics, bounded adversarial system disturbances, and state and input constraints. However, it involves solving a Partial Differential Equation (PDE), whose computational and memory complexity scales exponentially with respect to the state dimension, making its direct use on large-scale systems intractable. To overcome this challenge, neural approaches, such as DeepReach, have been used to synthesize reachable tubes and safety controllers for high-dimensional systems. However, verifying these neural reachable tubes remains challenging. In this work, we propose two different verification methods, based on robust scenario optimization and conformal prediction, to provide probabilistic safety guarantees for neural reachable tubes. Our methods allow a direct trade-off between resilience to outlier errors in the neural tube, which are inevitable in a learning-based approach, and the strength of the probabilistic safety guarantee. Furthermore, we show that split conformal prediction, a widely used method in the machine learning community for uncertainty quantification, reduces to a scenario-based approach, making the two methods equivalent not only for verification of neural reachable tubes but also more generally. To our knowledge, our proof is the first in the literature to show a strong relationship between the highly related but disparate fields of conformal prediction and scenario optimization. Finally, we propose an outlier-adjusted verification approach that harnesses information about the error distribution in neural reachable tubes to recover greater safe volumes. We demonstrate the efficacy of the proposed approaches for the high-dimensional problems of multi-vehicle collision avoidance and rocket landing with no-go zones.

Keywords: Probabilistic Safety Guarantees, Safety-Critical Learning, Neural Certificates, Hamilton-Jacobi Reachability Analysis, Scenario Optimization, Conformal Prediction

1. Introduction

It is important to design provably safe controllers for autonomous systems. Hamilton-Jacobi (HJ) reachability analysis provides a powerful framework to design such controllers for general nonlinear dynamical systems (Lygeros, 2004; Mitchell et al., 2005). In reachability analysis, safety is characterized by the system's *Backward Reachable Tube (BRT)*. This is the set of states from which trajectories will eventually reach a given target set despite the best control effort. Thus, if the target set represents undesirable states, the BRT represents unsafe states and should be avoided. Along with the BRT, reachability analysis provides a safety controller to keep the system outside the BRT.

Traditionally, the BRT computation in HJ reachability is formulated as an optimal control problem. The BRT can then be obtained as a sub-zero level solution of the corresponding value function. Obtaining the value function requires solving a partial differential equation (PDE) over a state-space

grid, resulting in an exponentially scaling computation complexity with the number of states (Bansal et al., 2017). To overcome this challenge, a variety of solutions have been proposed that trade off between the class of dynamics they can handle, the approximation quality of the BRT, and the required computation. These include specialized methods for linear and affine dynamics (Greenstreet and Mitchell, 1998; Frehse et al., 2011; Kurzhanski and Varaiya, 2000, 2002; Maidens et al., 2013; Girard, 2005; Althoff et al., 2010; Bak et al., 2019; Nilsson and Ozay, 2016), polynomial dynamics (Majumdar et al., 2014; Majumdar and Tedrake, 2017; Dreossi et al., 2016; Henrion and Korda, 2014), monotonic dynamics (Coogan and Arcak, 2015), and convex dynamics (Chow et al., 2017) (see Bansal et al. (2017); Bansal and Tomlin (2021) for a survey).

Owing to the success of deep learning, there has also been a surge of interest in approximating high-dimensional BRTs (Rubies-Royo et al., 2019; Fisac et al., 2019; Djeridane and Lygeros, 2006; Niarchos and Lygeros, 2006; Darbon et al., 2020) and optimal controllers (Onken et al., 2022) through deep neural networks (DNNs). Building upon this line of work, Bansal and Tomlin (2021) have proposed DeepReach – a toolbox that leverages recent advances in neural implicit representations and neural PDE solvers to compute a value function and a safety controller for high-dimensional systems. Compared to the aforementioned methods, DeepReach can handle general nonlinear dynamics, the presence of exogenous disturbances, as well as state and input constraints during the BRT computation. Consequently, methods for verifying neural reachable tubes have been proposed. For example, Lin and Bansal (2023) propose an iterative scenario-based method (Campi et al., 2009) to recover probabilistically safe reachable tubes from DeepReach solutions up to a desired confidence level and bound on violation rate. Unfortunately, the method does not allow an after-the-fact risk-return trade-off, and as a result, it is highly sensitive to outlier errors in the learned solutions. This can lead to highly conservative reachable tubes and a severe loss of recovery in the case of stringent safety requirements, as we demonstrate in our case studies.

In this work, we propose two different verification methods, one based on robust scenario optimization and the other based on conformal prediction, to provide probabilistic safety guarantees for neural reachable tubes. Both methods are resilient to the outlier errors in neural reachable tubes and automatically trade-off the strength of the probabilistic safety guarantees based on the outlier rate. The proposed methods can evaluate any candidate tube and are not restricted to a specific class of system dynamics or value functions. We further prove that these seemingly different verification methods naturally reduce to one another, providing a unifying viewpoint for uncertainty quantification (typical use case of conformal prediction) and error optimization (typical use case of scenario optimization) in neural reachable tubes. Based on these insights, we propose an outlier-adjusted verification approach that can recover a greater safe volume from a neural reachable tube by harnessing information about the distribution of error in the learned solution. To summarize, the key contributions of this paper are:

- probabilistic safety verification methods for neural reachable tubes that enable a direct trade-off between resilience and the probabilistic strength of safety,
- a proof that split conformal prediction reduces to a scenario-based approach in general, demonstrating a strong relationship between the two highly related but disparate fields,
- an outlier-adjusted verification approach that recovers greater safe volumes from tubes, and
- a demonstration of the proposed approaches for the high-dimensional problems of multi-vehicle collision avoidance and rocket landing with no-go zones.

2. Problem Setup

Consider a dynamical system with state $x \in X \subseteq \mathbb{R}^n$, control $u \in \mathcal{U}$, and dynamics $\dot{x} = f(x, u)$ governing how x evolves over time until a final time T. Let $\xi^u_{x,t}(\tau)$ denote the state achieved at time $\tau \in [t, T]$ by starting at initial state x and time t and applying control $u(\cdot)$ over $[t, \tau]$. Let $\mathcal{L}$ represent a target set that the agent wants to either reach (e.g. goal states) or avoid (e.g. obstacles).

Running example: Multi-Vehicle Collision Avoidance. Consider a 9D multi-vehicle collision avoidance system with 3 independent Dubins3D cars: Q_1, Q_2, Q_3. Q_i has position (p_{xi}, p_{yi}), heading θ_i, constant velocity v, and steering control $u_i \in [u_{\min}, u_{\max}]$. The dynamics of Q_i are: $\dot{p}_{xi} = v \cos\theta_i, \quad \dot{p}_{yi} = v \sin\theta_i, \quad \dot{\theta}_i = u_i$. $\mathcal{L}$ is the set of states where any of the vehicle pairs is in collision: $\mathcal{L} = \{x : \min\{d(Q_1, Q_2), d(Q_1, Q_3), d(Q_2, Q_3)\} \leq R\}$, where $d(Q_i, Q_j)$ is the distance between Q_i and Q_j. We set: $v = 0.6, \quad u_{\min} = -1.1, \quad u_{\max} = 1.1, \quad R = 0.25$.

In this setting, we are interested in computing the system's initial-time Backward Reachable Tube, which we denote as BRT. We define BRT as the set of all initial states in X from which the agent will eventually reach $\mathcal{L}$ within the time horizon $[0, T]$, despite best control efforts: BRT $= \{x : x \in X, \forall u(\cdot), \exists \tau \in [0, T], \xi^u_{x,0}(\tau) \in \mathcal{L}\}$. When $\mathcal{L}$ represents unsafe states for the system, as it does in our running example, staying outside of BRT is desirable. When $\mathcal{L}$ instead represents the states that the agent wants to reach, BRT is defined as the set of all initial states in X from which the agent, acting optimally, can eventually reach $\mathcal{L}$ within $[0, T]$. Thus, staying within BRT is desirable.

The above 9D system is intractable for traditional grid-based methods, motivating the use of DeepReach to learn a neural BRT. Our goal in this work is to recover an approximation of the safe set with probabilistic guarantees. Specifically, we want to find $\mathcal{S}$ such that $\mathbb{P}_{x \in \mathcal{S}}(x \in \text{BRT}) \leq \epsilon$ for some violation parameter $\epsilon \in (0, 1)$. When $\mathcal{L}$ represents goal states, we want $\mathbb{P}_{x \in \mathcal{S}}(x \in \text{BRT}^C) \leq \epsilon$.

3. Background: Hamilton-Jacobi Reachability, DeepReach, and Safety Verification

Here, we provide a quick overview of Hamilton-Jacobi reachability analysis, a specific toolbox, DeepReach, to compute high-dimensional neural reachable tubes, and an iterative scenario-based method for recovering probabilistically safe tubes from learning-based methods like DeepReach.

3.1. Hamilton-Jacobi (HJ) Reachability

In HJ reachability, computing BRT is formulated as an optimal control problem. We will explain it in the context of $\mathcal{L}$ being a set of undesirable states. In the end, we will comment on when $\mathcal{L}$ is a set of desirable states and refer interested readers to Bansal et al. (2017) for other cases.

We first define a target function $l(x)$ such that the sub-zero level of $l(x)$ yields $\mathcal{L}$: $\mathcal{L} = \{x : l(x) \leq 0\}$. $l(x)$ is commonly a signed distance function to $\mathcal{L}$. For example, we can choose $l(x) = \min\{d(Q_1, Q_2), d(Q_1, Q_3), d(Q_2, Q_3)\} - R$ for our running example in Section 2. Next, we define the cost function of a state corresponding to some policy $u(\cdot)$ to be the minimum of $l(x)$ over its trajectory: $J_{u(\cdot)}(x, t) = \min_{\tau \in [t,T]} l(\xi^u_{x,t}(\tau))$. Since the system wants to avoid $\mathcal{L}$, our goal is to maximize $J_{u(\cdot)}(x, t)$. Thus, the value function corresponding to this optimal control problem is:

$$V(x, t) = \sup_{u(\cdot)} J_{u(\cdot)}(x, t) \tag{1}$$

By defining our optimal control problem in this way, we can recover BRT using the value function. In particular, the value function being sub-zero implies that the target function is sub-zero somewhere along the optimal trajectory, or in other words, that the system has reached $\mathcal{L}$. Thus, BRT is given as the sub-zero level set of the value function at the initial time: BRT $= \{x : x \in$

$X, V(x,0) \leq 0\}$. The value function in Equation (1) can be computed using dynamic programming, resulting in the following final value Hamilton-Jacobi-Bellman Variational Inequality (HJB-VI): $\min\left\{D_t V(x,t) + H(x,t), l(x) - V(x,t)\right\} = 0$, with the terminal value function $V(x,T) = l(x)$. D_t and ∇ represent the time and spatial gradients of V. H is the Hamiltonian that encodes the role of dynamics and the optimal control: $H(x,t) = \max_u \langle \nabla V(x,t), f(x,u)\rangle$. The value function in Equation (1) induces the optimal safety controller: $u^*(x,t) = \arg\max_u \langle \nabla V(x,t), f(x,u)\rangle$. Intuitively, the safety controller aligns the system dynamics in the direction of the value function's gradient, thus steering the system towards higher-value states, i.e., away from $\mathcal{L}$.

We have just explained the case where $\mathcal{L}$ represents a set of undesirable states. When the system instead wants to reach $\mathcal{L}$, an infimum is used instead of a supremum in Equation (1). The control wants to reach $\mathcal{L}$, hence there is a minimum instead of a maximum in the Hamiltonian and optimal safety controller equations. See Bansal et al. (2017) for details on other reachability cases.

Traditionally, the value function is computed by solving the HJB-VI over a discretized grid in the state space. Unfortunately, doing so involves computation whose memory and time complexity scales exponentially with respect to the system dimension, making these methods practically intractable for high-dimensional systems, such as those beyond 5D. Fortunately, a deep learning approach, DeepReach, has been proposed to enable HJ reachability for high-dimensional systems.

3.2. DeepReach and an Iterative Scenario-Based Probabilistic Safety Verification Method

Instead of solving the HJB-VI over a grid, DeepReach (Bansal and Tomlin, 2021) learns a parameterized approximation of the value function using a sinusoidal deep neural network (DNN). Thus, memory and complexity requirements for training scale with the value function complexity rather than the grid resolution, allowing it to obtain BRTs for high-dimensional systems. DeepReach trains the DNN via self-supervision on the HJB-VI itself. Ultimately, it takes as input a state x and time t, and it outputs a learned value function $\tilde{V}(x,t)$. $\tilde{V}(x,t)$ also induces a corresponding safe policy $\tilde{\pi}(x,t)$, as well as a BRT (referred to as the neural reachable tube from hereon).

However, the neural reachable tube will only be as accurate as $\tilde{V}(x,t)$. To obtain a provably safe BRT, Lin and Bansal (2023) propose a uniform value correction bound which is defined, for the avoid case, as the maximum learned value of an unsafe state under the induced policy: $\delta_{\tilde{V},\tilde{\pi}} \coloneqq \max_{x\in X}\{\tilde{V}(x,0) : J_{\tilde{\pi}}(x,0) \leq 0\}$. The authors show that the super-$\delta_{\tilde{V},\tilde{\pi}}$ level set of $\tilde{V}(x,0)$ is provably safe under the policy $\tilde{\pi}(x,t)$. They also propose an iterative scenario-based probabilistic verification method for computing an approximation of $\delta_{\tilde{V},\tilde{\pi}}$ from finite random samples that satisfies a desired confidence level and violation rate. However, the method is sensitive to outlier errors in the neural reachable tube and can result in very conservative safe sets. Specifically, it does not provide safety assurances for safe sets with nonzero empirical safety violations.

In this work, we propose probabilistic safety verification methods that allow nonzero empirical safety violations at the cost of the probabilistic strength of safety. This enables a direct trade-off between resilience to outlier errors and the strength of the safety guarantee.

Remark 1 *Although we work with DeepReach solutions in particular for our problem setup, our proposed approaches can verify any general $\tilde{V}(x,t)$ and $\tilde{\pi}(x,t)$, regardless of whether DeepReach, a numerical PDE solver, or some other tool is used to obtain them.*

4. Robust Scenario-Based Probabilistic Safety Verification Method

Here, we propose a robust scenario-based probabilistic safety verification method for neural reachable tubes. The new method is a straightforward application of a scenario-based sampling-and-

discarding approach to chance-constrained optimization problems, which quantifies the trade-off between feasibility and performance of the optimal solution based on finite samples (Campi and Garatti, 2011). First, we explain the method when $\mathcal{L}$ represents undesirable states. In the end, we comment on when $\mathcal{L}$ represents desirable states.

Procedures: Let $\mathcal{S} \subseteq X$ be a neural safe set that is, in the avoid case, the *complement* of the neural reachable tube being verified. We propose to consider the superlevel sets of $\tilde{V}(x,0)$ as candidates for $\mathcal{S}$. Ideally, any super-δ level set of $\tilde{V}(x,0)$ for $\delta > 0$ should be a valid $\mathcal{S}$; however, due to learning errors, that might not be true in practice. In order to provide a probabilistic safety assurance for $\mathcal{S}$, we first sample N independent and identically distributed (i.i.d.) states $x_{1:N}$ from $\mathcal{S}$ according to some probability distribution $\mathbb{P}$ over $\mathcal{S}$. Since $\mathcal{S}$ is defined implicitly by $\tilde{V}(x,0)$, we use rejection sampling. We next compute the costs $J_{\tilde{\pi}}(x_i,0)$ for $i = 1,2,...,N$ by rolling out the system trajectory from x_i under $\tilde{\pi}(x,t)$. Let k refer to the number of "outliers" - samples that are empirically unsafe, i.e., $J_{\tilde{\pi}}(x_i,0) \leq 0$. Then the following theorem provides a probabilistic guarantee on the safety of the neural reachable tube and its complement, the neural safe set $\mathcal{S}$:

Theorem 2 (Robust Scenario-Based Probabilistic Safety Verification) *Select a safety violation parameter $\epsilon \in (0,1)$ and a confidence parameter $\beta \in (0,1)$ such that*

$$\sum_{i=0}^{k} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \beta \tag{2}$$

where k and N are as defined above. Then, with probability at least $1-\beta$ over the draws of the samples, the following holds:

$$\mathop{\mathbb{P}}_{x \in \mathcal{S}} (V(x,0) \leq 0) \leq \epsilon \tag{3}$$

All proofs can be found in the Appendix of the extended version of this article[1]. Disregarding the confidence parameter β for a moment, Theorem 2 states that the fraction of $\mathcal{S}$ that is unsafe is bounded above by the violation parameter ϵ, where ϵ is computed empirically using Equation (2) based on the outlier rate k encountered within N samples. ϵ is thus a reflection of the safety quality of $\mathcal{S}$, which degrades with the increase in the number of outliers k, as expected. This can also be seen for the running example in Figure 1 (the red curve). Overall, Theorem 2 allows us to compute probabilistic safety guarantees for any neural set $\mathcal{S}$ based on a finite number of samples. Subsequently, this result can be used to find some $\mathcal{S}$ for which ϵ is smaller than a desired threshold, as we discuss later in this section.

To interpret β, note that k is a random variable that depends on the randomly sampled $x_{1:N}$. It may be the case that we just happen to draw an unrepresentative sample, in which case the ϵ bound does not hold. β controls the probability of this adverse event happening, which regards the correctness of the probabilistic safety guarantee in Equation (3). Fortunately, β goes to 0 exponentially with N, so β can be chosen to be an extremely small value, such as 10^{-16}, when we sample large N. $1-\beta$ will then be so close to 1 that it does not have any practical importance.

We have just explained the robust scenario-based probabilistic safety verification method in the case where $\mathcal{L}$ represents undesirable states. When the system instead wants to reach $\mathcal{L}$, $\mathcal{S}$ will be a sublevel set instead of a superlevel set of the learned value function. The cost inequality should be flipped when computing k, and the value inequality should be flipped in Equation (3).

1. See https://sia-lab-git.github.io/Verification_of_Neural_Reachable_Tubes.pdf

4.1. Comparison of Robust and Iterative Scenario-Based Probabilistic Safety Verification

The key difference between the proposed robust scenario-based method and the iterative scenario-based method discussed in Section 3.2 is that the former can handle nonzero empirical safety violations k. This enables several crucial advantages that we demonstrate in Figures 1 and 2 for a solution learned by DeepReach on the multi-vehicle collision avoidance running example in Section 2. We have fixed the confidence parameter $\beta = 10^{-16}$ to be so close to 0 that it has no practical significance (β plays the same role in both methods).

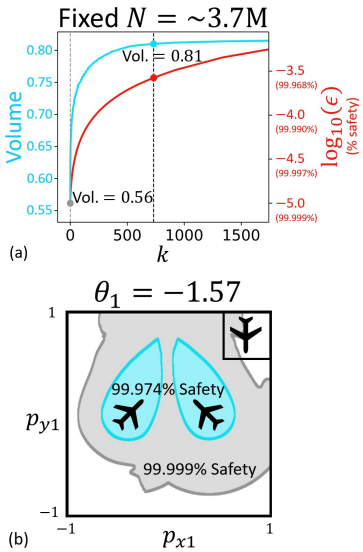

Figure 1: (Top) For a fixed simulation budget N, the cyan curve shows the number of empirical safety violations k for different learned volumes (different super-levels of $\tilde{V}(x, 0)$). The red curve shows the trade-off in safety strength ϵ (in log scale) for each k using the robust method. The grey point indicates the iterative method baseline. The robust method is able to provide safety assurances even for the volumes that have non-zero outliers. (Dashed black line) By a small decrease in safety level (from 99.999% to 99.974%) caused by outliers, we are able to significantly increase the assured safe volume from 0.56 to 0.81. (Bottom) Correspondingly, the safe set $\mathcal{S}$ increases greatly from the complement of the grey region to the complement of the blue region.

Firstly, for a fixed simulation budget N, the robust method allows one to trade off the probabilistic strength of safety (increasing ϵ) for resilience (increasing k). In other words, the method can verify any given neural safe set $\mathcal{S}$ in an outlier-robust fashion by automatically attenuating the level of safety assurance based on the number of empirical outliers (i.e., safety violations). The iterative method, in contrast, can only verify a region that is outlier-free. Consequently, the robust method enables one to engage in a trade-off if a large increase in safe set volume can be attained by a tolerable decrease in safety, as illustrated in Figure 1.

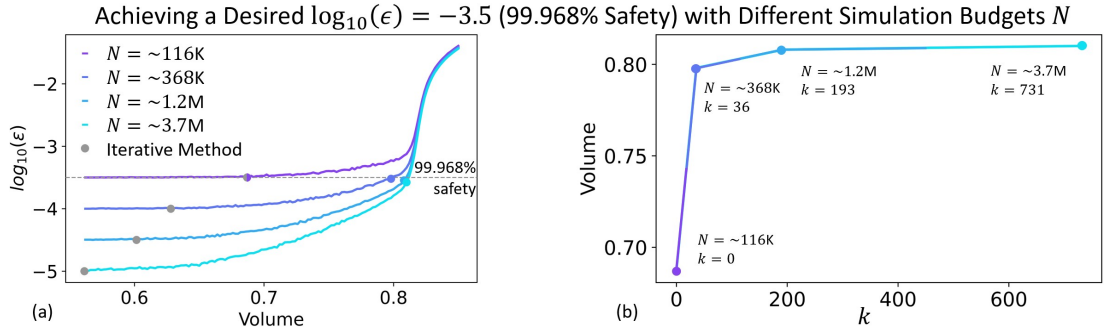

Figure 2: (Left) Computing the safety strength ϵ across different volumes (different super-levels of $\tilde{V}(x, 0)$) for different simulation budgets N using the robust method. The grey points indicate the iterative method baselines. (Right) As we increase N, the largest volume achieving the desired 99.968% safety using the robust method increases up to a limit.

Secondly, by allowing nonzero safety violations k, the robust method provides stronger safety assurances for a *fixed* volume with increment in the simulation budget N, as long as the outlier rate

does not grow substantially with N. Thus, with more simulation effort, significantly larger volumes can be attained *for a desired safety strength* ϵ as shown in Figure 2. Incrementing N in the iterative method, on the other hand, will only correspond to verifying *smaller* volumes at a *stronger* ϵ. It cannot verify larger volumes for a fixed ϵ, because empirical safety violations will be introduced. Figure 2 shows how the robust method (curves) adds a new degree of freedom for computing safety assurances compared to the iterative method (grey points).

5. Conformal Probabilistic Safety Verification Method

We now propose a *conformal* probabilistic safety verification method for neural reachable tubes which is intended to be the direct analogue of the *robust scenario-based* method in Section 4. The method is a straightforward application of split conformal prediction, a widely used method in the machine learning community for uncertainty quantification (Angelopoulos and Bates, 2023).

Using the same procedures as described in Section 4, split conformal prediction can be used instead of robust scenario optimization to provide a probabilistic guarantee on the safety of the neural reachable tube and its complement, the neural safe set $\mathcal{S}$:

Theorem 3 (Conformal Probabilistic Safety Verification) *Let the number of outliers k and the number of samples N be as defined in the procedures in Section 4, then:*

$$\mathop{\mathbb{P}}_{x\in\mathcal{S}}\left(J_{\tilde{\pi}}(x,0)>0\right)\sim \text{Beta}(N-k,k+1) \tag{4}$$

Theorem 3 can be established via a straightforward application of conformal prediction with $-J_{\tilde{\pi}}(x,0)$ as the scoring function. The proof is in the Appendix of the extended version of this article[1]. The above theorem states that the fraction of $\mathcal{S}$ that is safe is distributed according to the Beta distribution with shape parameters $N-k$ and $k+1$. Intuitively, the mass in the distribution shifts towards 0 as k increases for a fixed N, implying that it is more likely that a smaller fraction of $\mathcal{S}$ is safe, as expected. For a fixed ratio $N:k$, N controls how concentrated the mass is around the mean; i.e., for larger sample sizes N, we can more confidently determine the fraction of $\mathcal{S}$ that is safe.

To better understand Theorem 3, we show in Figure 3 the Beta distribution of $\mathop{\mathbb{P}}_{x\in\mathcal{S}}\left(J_{\tilde{\pi}}(x,0)>0\right)$ for a solution learned by DeepReach on the multi-vehicle collision avoidance running example in Section 2, for which $k=731$ outliers are found from $N=3684118$ samples.

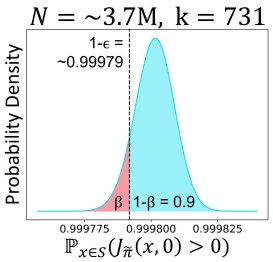

Figure 3: The Beta distribution of $\mathop{\mathbb{P}}_{x\in\mathcal{S}}\left(J_{\tilde{\pi}}(x,0)>0\right)$ when $k=731$ outliers are found from $N=3684118$ samples. (Dashed black line) For an example choice of confidence $1-\beta=0.9$ (shaded blue), we can lower-bound the fraction of $\mathcal{S}$ which is safe with at least $1-\epsilon=0.99979$ (99.979%) confidence.

Remark 4 *The mean of the Beta distribution in Equation (4) is given as $\frac{N-k}{N+1}$, which is roughly the fraction of the empirically safe samples. One can immediately derive that the safety probability of $\mathcal{S}$, marginalized over the sampled "calibration" states, is given as: $\mathop{\mathbb{P}}_{(x_{1:N},x)\in\mathcal{S}}\left(J_{\tilde{\pi}}(x,0)>0\right)\geq\frac{N-k}{N+1}$, which precisely resembles the most commonly used **coverage property** of split conformal prediction.*

Even though Theorem 3 provides the distribution of the safety level, when we compute safety assurances in practice, it is often desirable to know a lower-bound on the safety level with at least some desired confidence. This corresponds to choosing a lower-bound whose accumulated probability mass is smaller than some confidence parameter β (shaded red in Figure 3). The following lemma formalizes this by using the CDF of the Beta distribution in Theorem 3.

Lemma 5 (Conformal Probabilistic Safety Verification) *Select a safety violation parameter $\epsilon \in (0,1)$ and a confidence parameter $\beta \in (0,1)$ such that*

$$\sum_{i=0}^{k} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \beta \tag{5}$$

where k and N are as defined above. Then, with probability at least $1-\beta$ over the draws of the samples, the following holds:

$$\mathop{\mathbb{P}}_{x \in \mathcal{S}} \left(V(x,0) \leq 0 \right) \leq \epsilon \tag{6}$$

Lemma 5 is, in fact, precisely the same result as obtained by Theorem 2 using robust scenario optimization. This is no coincidence, as one can show that split conform prediction more generally reduces to a robust scenario-optimization problem.

Remark 6 *In general, a split conformal prediction problem can be reduced to a robust scenario-optimization problem. This is proven in the Appendix of the extended version of this article.*[1]

Due to the equivalence between conformal method and robust scenario-based methods, the analysis in Section 4 holds here as well. More generally, we hope that this insight will lead to future research into further investigating the close relationship between the two methods.

6. Outlier-Adjusted Probabilistic Safety Verification Approach

The verification methods in Sections 4 and 5 are limited by the quality of the neural reachable tube. Although they can account for outliers, the computed safety level can be low if the outlier rate is high. This can lead to significant losses in the safe volume, as demonstrated in Sections 6.2 and 6.3.

To address this issue, we propose an outlier-adjusted approach that can recover a larger safe volume for any desired ϵ. Note that in the verification methods, the key quantity which determines ϵ is the number of safety violations k. This corresponds to the number of samples x_i which are marked safe by membership in $\mathcal{S}$, i.e., $\tilde{V}(x_i,0) \geq \delta$, but are not guaranteed to be safe, i.e., $J_{\tilde{\pi}}(x_i,0) \leq 0$. It is easy to see that the best we can do to simultaneously minimize k and maximize volume is to compute $\mathcal{S}$ as the super-δ level set of the induced *cost* function $J_{\tilde{\pi}}(x,0)$. For example, the largest possible $\mathcal{S}$ that is guaranteed to be violation-free is precisely the super-zero level set of $J_{\tilde{\pi}}(x,0)$. Thus, our overall approach will be to refine $\tilde{V}(x,0)$ so that it more accurately reflects $J_{\tilde{\pi}}(x,0)$.

Modeling $J_{\tilde{\pi}}(x,0)$ can be formulated as a supervised learning problem, since we can sample a state x_i and compute its cost $J_{\tilde{\pi}}(x_i,0)$ in simulation. We learn an approximation $\tilde{J}_{\tilde{\pi}}(x,0)$ by retraining $\tilde{V}(x,0)$ on a training dataset $\mathcal{T}$ of n samples, $\mathcal{T} = (x_1, J_{\tilde{\pi}}(x_1,0)), ..., (x_n, J_{\tilde{\pi}}(x_n,0))$. Specifically, we use the *weighted* MSE loss $\frac{1}{n}\sum_{i=1}^{n} w_i (\tilde{V}(x_i,0) - J_{\tilde{\pi}}(x_i,0))^2$, where $w_i = w$ if the error is conservative $\left(\tilde{V}(x_i,0) < J_{\tilde{\pi}}(x_i,0)\right)$, otherwise $w_i = 1$. We introduce w as a hyperparameter to underweight conservative errors because in the end, we are concerned with recovering

larger *safe* volumes. Thus, selecting a small w allows us to focus on reducing *optimistic* errors $\left(\tilde{V}(x_i, 0) > J_{\tilde{\pi}}(x_i, 0)\right)$ which are more safety-critical and correspond to outlier safety violations.

To avoid overfitting, we select the training checkpoint that performs best on a validation dataset $\mathcal{V}$. The validation metric we use is the maximum learned cost of an empirically unsafe state: $\max_{x \in \mathcal{V}} \{\tilde{J}_{\tilde{\pi}}(x, 0) : J_{\tilde{\pi}}(x, 0) \leq 0\}$, which one can think of as a proxy for the recoverable safe volume. We demonstrate the efficacy of the proposed outlier-adjusted approach for the high-dimensional systems of multi-vehicle collision avoidance and rocket landing with no-go zones. For all case studies, we set $w = 10^{-3}$ during retraining, fix the confidence parameter $\beta = 10^{-16}$ and find a safe volume that satisfies $\epsilon \leq 10^{-4}$ (99.990% safety) using the robust method in Section 4.

6.1. Multi-Vehicle Collision Avoidance

In Figure 4, we compare our outlier-adjusted approach (blue) to the baseline (grey) for a DeepReach solution trained on the multi-vehicle collision avoidance running example in Section 2. A 2.3% increase in the safe volume is attained, shown by the tightened BRT. Note that the largest visual difference in the BRT is where the third vehicle is between the two others; intuitively, the safety in this region is likely more difficult to model by the baseline approach.

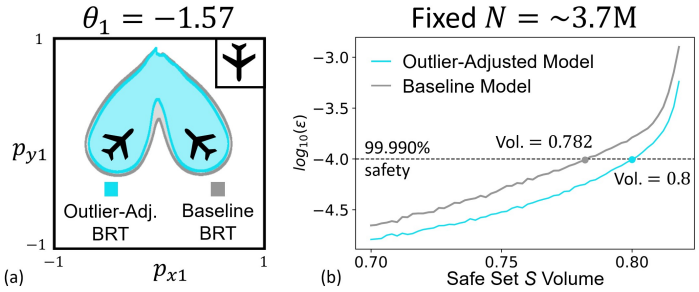

Figure 4: Multi-Vehicle Collision Avoidance: outlier-adjusted (blue) and baseline (grey) results. (Left) Slice of the neural BRTs achieving $\epsilon = 10^{-4}$ (99.990% safety). (Right) The outlier-adjusted approach increases the safe volume from 0.782 to 0.8 (2.3% increase).

6.2. Rocket Landing

We now apply our approach to a 6D rocket landing system with position (p_x, p_y), heading θ, velocity (v_x, v_y), angular velocity ω, and torque controls $\tau_1, \tau_2 \in [-250, 250]$. The dynamics are: $\dot{p_x} = v_x$, $\dot{p_y} = v_y$, $\dot{\theta} = \omega$, $\dot{\omega} = 0.3\tau_1$, $\dot{v_x} = \tau_1 \cos\theta - \tau_2 \sin\theta$, $\dot{v_y} = \tau_1 \sin\theta + \tau_2 \cos\theta - g$, where $g = 9.81$ is acceleration due to gravity. The target set is the set of states where the rocket reaches a rectangular landing zone of side length 20m centered at the origin: $\mathcal{L} = \{x : |p_x| < 20.0, p_y < 20.0\}$. Note that we want to *reach* $\mathcal{L}$, so the BRT now represents the safe set. Results are shown in Figure 5. Interestingly, a large 9.58% increase in the volume of the safe set is recovered using the proposed approach, particularly near the lower-left part of the state space. Further investigation reveals that the trajectories starting from these states exit the training regime south. This highlights a general limitation of computing the value function over a constrained state space where information is propagated via dynamic programming, which affects both learning-based methods and traditional grid-based methods. Nevertheless, in this case, the relative order of the value function levels is still preserved, leading to a high quality safe policy and recovery of a larger safe volume.

6.3. Rocket Landing with No-Go Zones

We now consider the rocket landing problem in a constrained airspace where we have no-go zones of height 100m and width 10m to the left of the landing zone and where altitude is below the landing zone. Safety in this case takes the form of a reach-avoid set - the rocket needs to reach the

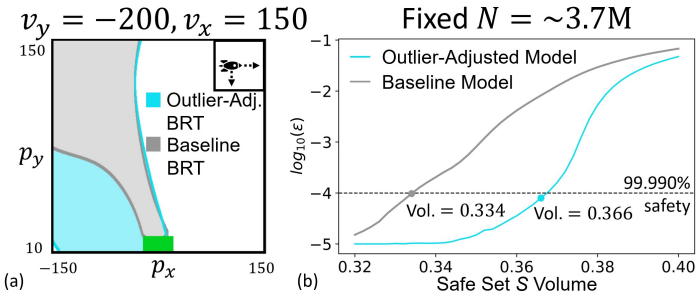

Figure 5: Rocket Landing: outlier-adjusted (blue) and baseline (grey) results. (Left) Slice of the neural BRTs achieving $\epsilon = 10^{-4}$ (99.990% safety). (Right) The outlier-adjusted approach increases the safe volume from 0.334 to 0.366 (9.58% increase).

landing zone while avoiding the no-go zones. An analogous HJI-VI to the one in Section 3.1 can be derived for this case, whose solution can be computed using DeepReach. However, since reach-avoid problems are more complex than just the reach or avoid problem, the DeepReach solution results in a poor safety volume. In fact, *no* safe volume can be recovered with the desired safety level of $\epsilon \leq 10^{-4}$. In contrast, we can recover a sizable safe volume using the outlier-adjusted approach, as shown in Figure 6. These examples highlight the utility of the proposed approach.

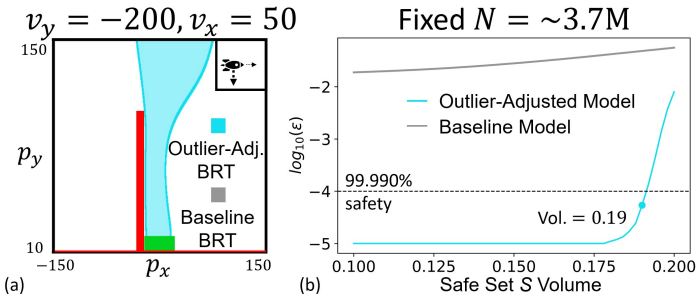

Figure 6: Rocket Landing with No-Go Zones: outlier-adjusted (blue) and baseline (grey) results. (Left) Slice of the neural BRTs achieving $\epsilon = 10^{-4}$ (99.990% safety). (Right) The outlier-adjusted approach increases the safe volume from 0 to 0.19.

7. Discussion and Future Work

In this work, we propose two different verification methods, based on robust scenario optimization and conformal prediction, to provide probabilistic safety guarantees for neural reachable tubes. Our methods allow a direct trade-off between resilience to outlier errors in the neural tube, which are inevitable in a learning-based approach, and the strength of the probabilistic safety guarantee. Furthermore, we show that split conformal prediction, a widely used method in the machine learning community for uncertainty quantification, reduces to a scenario-based approach, making the two methods equivalent not only for verification of neural reachable tubes but also more generally. We hope that our proof will lead to future insights into the close relationship between the highly related but disparate fields of conformal prediction and scenario optimization. Finally, we propose an outlier-adjusted verification approach that harnesses information about the error distribution in neural reachable tubes to recover greater safe volumes. We demonstrate the efficacy of the proposed approaches for the high-dimensional problems of multi-vehicle collision avoidance and rocket landing with no-go zones. Altogether, these are important steps toward using learning-based reachability methods to compute safety assurances for high-dimensional systems in the real world.

In the future, we will explore how the key idea of the outlier-adjusted verification approach, using cost labels as a supervised learning signal, can be used to enhance the accuracy of learning-based reachability methods like DeepReach. Other directions include providing safety assurances in the presence of worst-case disturbances and in real-time for tubes that are generated online.

Acknowledgments

This work is supported in part by a NASA Space Technology Graduate Research Opportunity, the NVIDIA Academic Hardware Grant Program, the NSF CAREER Program under award 2240163, and the DARPA ANSR program.

References

Matthias Althoff, Olaf Stursberg, and Martin Buss. Computing reachable sets of hybrid systems using a combination of zonotopes and polytopes. *Nonlinear analysis: hybrid systems*, 4(2):233–249, 2010.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023. ISSN 1935-8237. doi: 10.1561/2200000101. URL http://dx.doi.org/10.1561/2200000101.

Stanley Bak, Hoang-Dung Tran, and Taylor T Johnson. Numerical verification of affine systems with up to a billion dimensions. In *International Conference on Hybrid Systems: Computation and Control*, pages 23–32, 2019.

Somil Bansal and Claire J Tomlin. DeepReach: A deep learning approach to high-dimensional reachability. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J Tomlin. Hamilton-Jacobi Reachability: A brief overview and recent advances. In *IEEE Conference on Decision and Control (CDC)*, 2017.

M. C. Campi and S. Garatti. A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. *Journal of Optimization Theory and Applications*, 2011.

M. C. Campi, S. Garatti, and M. Prandini. The scenario approach for systems and control design. *Annual Reviews in Control*, 2009.

Yat Tin Chow, Jérôme Darbon, Stanley Osher, and Wotao Yin. Algorithm for overcoming the curse of dimensionality for time-dependent non-convex hamilton–jacobi equations arising from optimal control and differential games problems. *Journal of Scientific Computing*, 73(2-3):617–643, 2017.

Samuel Coogan and Murat Arcak. Efficient finite abstraction of mixed monotone systems. In *Proceedings of the 18th International Conference on Hybrid Systems: Computation and Control*, pages 58–67, 2015.

Jerome Darbon, Gabriel P Langlois, and Tingwei Meng. Overcoming the curse of dimensionality for some hamilton–jacobi partial differential equations via neural network architectures. *Research in the Mathematical Sciences*, 7(3):1–50, 2020.

Badis Djeridane and John Lygeros. Neural approximation of pde solutions: An application to reachability computations. In *Conference on Decision and Control*, pages 3034–3039, 2006.

Tommaso Dreossi, Thao Dang, and Carla Piazza. Parallelotope bundles for polynomial reachability. In *International Conference on Hybrid Systems: Computation and Control*, 2016.

Jaime F. Fisac, Neil F. Lugovoy, Vicenç Rubies-Royo, Shromona Ghosh, and Claire J. Tomlin. Bridging Hamilton-Jacobi Safety Analysis and Reinforcement Learning. *International Conference on Robotics and Automation*, 2019.

G. Frehse, C. Le Guernic, A. Donzé, S. Cotton, R. Ray, O. Lebeltel, R. Ripado, A. Girard, T. Dang, and O. Maler. SpaceEx: Scalable verification of hybrid systems. In *International Conference Computer Aided Verification*, 2011.

Antoine Girard. Reachability of uncertain linear systems using zonotopes. In *International Workshop on Hybrid Systems: Computation and Control*, pages 291–305, 2005.

Mark R. Greenstreet and Ian Mitchell. Integrating projections. In Thomas A. Henzinger and Shankar Sastry, editors, *Hybrid Systems: Computation and Control*, pages 159–174, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg. ISBN 978-3-540-69754-1.

D. Henrion and M. Korda. Convex computation of the region of attraction of polynomial control systems. *IEEE Transactions on Automatic Control*, 59(2):297–312, 2014.

Alexander Kurzhanski and Pravin Varaiya. On ellipsoidal techniques for reachability analysis. part ii: Internal approximations box-valued constraints. *Optimization Methods and Software*, 17:207–237, 01 2002. doi: 10.1080/1055678021000012435.

Alexander B Kurzhanski and Pravin Varaiya. Ellipsoidal techniques for reachability analysis: internal approximation. *Systems & Control Letters*, 2000.

Albert Lin and Somil Bansal. Generating formal safety assurances for high-dimensional reachability. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10525–10531. IEEE, 2023.

John Lygeros. On reachability and minimum cost optimal control. *Automatica*, 40(6):917–927, 2004.

John N Maidens, Shahab Kaynama, Ian M Mitchell, Meeko MK Oishi, and Guy A Dumont. Lagrangian methods for approximating the viability kernel in high-dimensional systems. *Automatica*, 2013.

A. Majumdar and R. Tedrake. Funnel libraries for real-time robust feedback motion planning. *The International Journal of Robotics Research*, 36(8):947–982, 2017.

Anirudha Majumdar, Ram Vasudevan, Mark M. Tobenkin, and Russ Tedrake. Convex optimization of nonlinear feedback controllers via occupation measures. *The International Journal of Robotics Research*, 33(9):1209–1230, 2014. doi: 10.1177/0278364914528059. URL https://doi.org/10.1177/0278364914528059.

Ian Mitchell, Alex Bayen, and Claire J. Tomlin. A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games. *IEEE Transactions on Automatic Control (TAC)*, 50(7):947–957, 2005.

KN Niarchos and John Lygeros. A neural approximation to continuous time reachability computations. In *Conference on Decision and Control*, pages 6313–6318, 2006.

Petter Nilsson and Necmiye Ozay. Synthesis of separable controlled invariant sets for modular local control design. In *American Control Conference*, pages 5656–5663, 2016.

Derek Onken, Levon Nurbekyan, Xingjian Li, Samy Wu Fung, Stanley Osher, and Lars Ruthotto. A neural network approach for high-dimensional optimal control applied to multiagent path finding. *IEEE Transactions on Control Systems Technology*, 2022.

Vicenç Rubies-Royo, David Fridovich-Keil, Sylvia Herbert, and Claire J Tomlin. A classification-based approach for approximate reachability. In *International Conference on Robotics and Automation*, pages 7697–7704. IEEE, 2019.