# Tight Certified Robustness via Min-Max Representations of ReLU Neural Networks

Brendon G. Anderson    Samuel Pfrommer    Somayeh Sojoudi

*Abstract*— **The reliable deployment of neural networks in control systems requires rigorous robustness guarantees. In this paper, we obtain *tight* robustness certificates over convex attack sets for min-max representations of ReLU neural networks by developing a convex reformulation of the nonconvex certification problem. This is done by "lifting" the problem to an infinite-dimensional optimization over probability measures, leveraging recent results in distributionally robust optimization to solve for an optimal discrete distribution, and proving that solutions of the original nonconvex problem are generated by the discrete distribution under mild boundedness, nonredundancy, and Slater conditions. As a consequence, optimal (worst-case) attacks against the model may be solved for *exactly*. This contrasts prior state-of-the-art that either requires expensive branch-and-bound schemes or loose relaxation techniques. Experiments on robust control and MNIST image classification examples highlight the benefits of our approach.**

## I. Introduction

Neural networks are rapidly being deployed in control systems as a means to efficiently model nonlinear systems [1], controllers [2], and reinforcement learning policies [3]. However, the performance of neural networks can be extremely sensitive to small fluctuations in their input data [4]. For example, [5], [6] show that image classification models can be fooled into misclassifying vehicle traffic signs when subject to digital or physical adversarial attacks, i.e., human-imperceptible data perturbations designed to cause failure. This unreliable behavior is directly at odds with the robustness guarantees required in safety-critical control settings such as autonomous driving [7].

In light of these sensitivities, much effort has been placed on developing methods to rigorously certify the robustness of neural networks, with a large emphasis on models using the popular ReLU activation function. However, certifying a neural network's robustness generally amounts to solving an intractable nonconvex optimization problem [8]. Three major lines of work have focused on overcoming this intractability: convex relaxations, Lipschitz bounding, and branch-and-bound methods (all discussed further in Section I-A).

In this paper, we utilize an alternative representation of ReLU neural networks as a means to efficiently compute tight robustness certificates using convex optimization (and hence in polynomial time). As a consequence, we are able to exactly compute optimal (worst-case) attacks, which is generally not possible using the popular local search-based attack methods such as projected gradient descent [9] and the Carlini-Wagner attack [10].

### A. Related Works

*1) Robustness Certification:* Certifying the robustness of a model amounts to solving the nonconvex optimization $\inf_{x\in X} g(x)$, where $X$ is a set of possible inputs or attacks (i.e., the "threat model"), and $g(x)$ is either the model output at an input $x$, or some linear transformation of the model output (e.g., a classifier's margin between two classes).

Convex relaxations work by optimizing over a convex outer-approximation of the set $g(X)$ of possible outputs. Popular relaxations involve linear bounding and programming [11], [12], and semidefinite programming [13], [14], which constitutes a line of increasingly accurate yet computationally complex relaxations. Convex relaxation-based certificates remain loose in general, and their looseness has been shown to increase with model size [15].

The Lipschitz constant of a model provides a certified bound on how much the model output may change given some change in its input. Thus, bounds on the Lipschitz constant can yield efficient robustness certificates [16]. A number of works are devoted to computing Lipschitz bounds, but it has proven difficult to obtain tight enough bounds to grant meaningful certificates [16], [17], [18], [19].

Mixed-integer programming and branch-and-bound have also been applied to robustness certification for ReLU neural networks [20], [21], [22]. In contrast to convex relaxations and Lipschitz bounding, these methods are capable of obtaining tight certificates if they are run to convergence, but this incurs exponential computational complexity, preventing them from scaling to practically-sized models [22]. Some methods allow for early termination of their optimizations to yield more efficient, yet loose certificates [22].

*2) Representations of ReLU Neural Networks:* ReLU neural networks are defined by compositions $g = \mathcal{A}_L \circ \sigma \circ \cdots \circ \sigma \circ \mathcal{A}_1$ with affine functions $\mathcal{A}_l$ and elementwise activation functions $\sigma = \mathrm{ReLU}\colon x \mapsto \max\{0, x\}$. The most prevalent alternative representation of such a model is as a piecewise linear function, i.e., a finite polyhedral partition of $\mathbb{R}^d$ with associated affine functions that agree with $g$ on each polyhedron [23], [24]. Another representation is as a rational function when working with tropical algebra, where addition $\oplus$ and multiplication $\otimes$ are defined by $x \oplus y = \max\{x, y\}$ and $x \oplus y = x + y$ [25]. Finally, min-max representations—discussed in Section II—have recently been introduced, where $g$ is expressed as the pointwise minimum of pointwise maxima of affine functions. These works restrict their focus to showcasing the impressive approximation capabilities of ReLU models and their alternative representations.

The authors are with the University of California, Berkeley. Emails: {bganderson,sam.pfrommer,sojoudi}@berkeley.edu.

This work was supported by grants from ONR and NSF.

### B. Contributions

1) We show that ReLU neural networks admit min-max representations and hence such representations are universal function approximators.
2) By lifting the certification to an infinite-dimensional problem over probability measures, we prove that, under mild boundedness, nonredundancy, and Slater conditions, *exact* solutions to the original nonconvex problem are efficiently obtained for min-max representations via reduction to a tractable finite-dimensional convex optimization problem.
3) Experiments on robust control and MNIST image classification examples demonstrate the effectiveness of our approach.

To the best of our knowledge, our work is the first to grant *tight* robustness certificates in polynomial time amongst those considering general ReLU neural networks and their alternative representations.[1]

### C. Organization

In Section II, we introduce and analyze the min-max representation of ReLU neural networks. We develop our tight robustness certificates in Section III. Experiments illustrating the effectiveness of our approach are given in Section IV, and concluding remarks are made in Section V.

### D. Notations

The sets of natural, real, nonnegative real, and extended real numbers are denoted by $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_+$, and $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ respectively. Throughout, we let $\mathcal{I}, \mathcal{J}, \mathcal{K} \subseteq \mathbb{N}$ denote index sets $\{1, \dots, m\}$, $\{1, \dots, n\}$, and $\{1, \dots, p\}$, respectively. The cardinality, convex hull, conic hull, and relative interior of a subset $X$ of $\mathbb{R}^d$ are denoted by $|X|$, $\operatorname{conv}(X)$, $\operatorname{cone}(X)$, and $\operatorname{ri}(X)$, respectively. Furthermore, we define $\mathcal{B}(X)$ to be the Borel $\sigma$-algebra on $X$. We denote the set of probability measures on the measurable space $(X, \mathcal{B}(X))$ by $\mathcal{P}(X)$. For $x \in \mathbb{R}^d$, the Dirac measure centered at $x$ is denoted by $\delta_x$, which we recall is the probability measure defined by $\delta_x(A) = 0$ if $x \notin A$ and $\delta_x(A) = 1$ if $x \in A$ for all $A \in \mathcal{B}(X)$. The set of all Dirac measures with center in $X$ is defined to be $\mathcal{D}(X) = \{\mu \in \mathcal{P}(X) : \mu = \delta_x \text{ for some } x \in X\}$. The set of continuous functions from $\mathbb{R}^d$ into $\mathbb{R}$ is denoted by $C(\mathbb{R}^d, \mathbb{R})$. The effective domain of a function $f \colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is defined to be the set $\operatorname{dom}(f) = \{x \in \mathbb{R}^d : f(x) < \infty\}$. If $f$ is Borel measurable and $\mu$ is a probability measure on $(X, \mathcal{B}(X))$, then we denote the expected value of $f$ with respect to $\mu$ by $\mathbb{E}_{x \sim \mu} f(x) = \int_X f(x) d\mu(x)$. If $f$ is convex, the subdifferential of $f$ at $x$ is denoted by $\partial f(x)$. Throughout, we let $\|\cdot\|$ denote an arbitrary norm on $\mathbb{R}^d$, and we denote its dual norm by $\|\cdot\|_*$.

## II. Min-Max Affine Functions

In this section, we formally define min-max affine functions, discuss works related to these functions, and show that every ReLU neural network admits such a representation.

**Definition 1.** A function $g \colon \mathbb{R}^d \to \mathbb{R}$ is a *min-max affine function* if there exist $\mathcal{I}, \mathcal{J}_1, \dots, \mathcal{J}_{|\mathcal{I}|} \subseteq \mathbb{N}$ and associated $a_{ij} \in \mathbb{R}^d$, $b_{ij} \in \mathbb{R}$ such that $g(x) = \min_{i \in \mathcal{I}} \max_{j \in \mathcal{J}_i} (a_{ij}^\top x + b_{ij})$ for all $x \in \mathbb{R}^d$. In this case, the function $x \mapsto \min_{i \in \mathcal{I}} \max_{j \in \mathcal{J}_i} (a_{ij}^\top x + b_{ij})$ is called the *min-max representation* of $g$.

The class of all min-max affine functions on $\mathbb{R}^d$ is denoted by $\mathcal{G}$. Notice that $g \in \mathcal{G}$ is the pointwise minimum of $m = |\mathcal{I}|$ convex functions $g_i \colon x \mapsto \max_{j \in \mathcal{J}_i} (a_{ij}^\top x + b_{ij})$, and is therefore nonconvex in general. Without loss of generality, we henceforth assume that, for every $g \in \mathcal{G}$, there exists some $\mathcal{J} \subseteq \mathbb{N}$ with $n = |\mathcal{J}|$ such that the min-max representation of $g$ satisfies $\mathcal{J}_i = \mathcal{J}$ for all $i \in \mathcal{I}$.[2]

**Related Works on Min-Max Affine Functions.** In the mathematics literature, min-max affine functions are also termed lattice polynomials [28]. The work [29] shows that piecewise linear activation functions can be written in min-max affine form, and that neural networks learned with such representations perform highly in image classification tasks. The works [30], [31] study the theoretical and algorithmic aspects of training min-max affine functions to separate data, and show that separating $\{x_1, \dots, x_p\} \subseteq \mathbb{R}^d$ from $\{y_1, \dots, y_q\} \subseteq \mathbb{R}^d$ requires no more that $pq$ affine components. The authors of [32] use min-max representations of neural networks to characterize the training optimization landscape. The conversion of ReLU neural networks into min-max affine form is characterized in [33, Theorem 4.15]. An algorithm for nonlinear system identification using min-max affine functions is developed in [34]. Finally, min-max affine functions have been used as consistent statistical estimators, termed "Riesz estimators," in the mathematical economics literature [35]. To the best of our knowledge, our work is the first to exploit min-max representations for purposes of robustness certification.

We now proceed with analyzing the representation power min-max affine functions. Let $\mathcal{F}$ be the class of all ReLU neural networks on $\mathbb{R}^d$. The following theorem shows that every ReLU neural network can be represented as a min-max affine function, and therefore min-max affine functions are universal function approximators.

**Theorem 1.** *For every $f \in \mathcal{F}$, there exist $\mathcal{I}, \mathcal{J} \subseteq \mathbb{N}$ and $(a_{ij}, b_{ij}) \in \mathbb{R}^d \times \mathbb{R}$ for $i \in \mathcal{I}$, $j \in \mathcal{J}$ such that*

$$f(x) = \min_{i \in \mathcal{I}} \max_{j \in \mathcal{J}} (a_{ij}^\top x + b_{ij}) \text{ for all } x \in \mathbb{R}^d. \tag{1}$$

[1] See [26] for a polynomial time solution to the special case of 2-layer ReLU models.

[2] This is without loss of generality, since the value $g(x)$ does not change upon appending affine global underestimators of the convex function $g_i \colon x \mapsto \max_{j \in \mathcal{J}_i} (a_{ij}^\top x + b_{ij})$ to the set of affine components $x \mapsto a_{ij}^\top x + b_{ij}$ of $g$. In other words, $\min_{i \in \mathcal{I}} \max_{j \in \mathcal{J}_i} (a_{ij}^\top x + b_{ij}) = \min_{i \in \mathcal{I}} \max_{j \in \mathcal{J}} (a_{ij}^\top x + b_{ij})$ if one defines $\mathcal{J} = \{1, \dots, n\}$ with $n = \max_{i \in \mathcal{I}} |\mathcal{J}_i|$ and $a_{ij} = v_i$, $b_{ij} = g_i(0)$ for $j \in \mathcal{J} \setminus \mathcal{J}_i$, for all $i \in \mathcal{I}$, where $v_i \in \mathbb{R}^d$ is a subgradient of $g_i$ at 0 (which exists by [27, Theorem 23.4]).

*Hence, the class $\mathcal{G}$ of min-max affine functions is dense in $C(\mathbb{R}^d, \mathbb{R})$ with respect to the topology of uniform convergence on compact sets.*

*Proof.* Every $f \in \mathcal{F}$ is piecewise affine, i.e., there is a finite collection $\mathcal{Q}$ of closed subsets of $\mathbb{R}^d$ such that $\mathbb{R}^d = \bigcup_{Q \in \mathcal{Q}} Q$ and $f$ is affine on every $Q \in \mathcal{Q}$. Hence, by [36, Theorem 4.1], there exist $\mathcal{I}, \mathcal{J} \subseteq \mathbb{N}$ and $(a_{ij}, b_{ij}) \in \mathbb{R}^d \times \mathbb{R}$ for $i \in \mathcal{I},\ j \in \mathcal{J}$ such that (1) holds. Thus, since $\mathcal{F} \subseteq \mathcal{G}$ and $\mathcal{F}$ is dense in $C(\mathbb{R}^d, \mathbb{R})$ with respect to the topology of uniform convergence on compact sets [37, Theorem 3.1], it holds that $\mathcal{G}$ is dense in $C(\mathbb{R}^d, \mathbb{R})$ in the same sense. □

## III. Theoretical Robustness Certificates

In this section, we develop our theoretical robustness certificates. Consider a model $g \colon \mathbb{R}^d \to \overline{\mathbb{R}}$, which may, for example, represent the output of a scalar-valued controller or the confidence of a binary classifier $f \colon \mathbb{R}^d \to \{1, 2\}$ defined by $f(x) = 1$ if $g(x) \geq 0$ and $f(x) = 2$ if $g(x) < 0$. We consider the asymmetric robustness setting introduced in [38], where nonnegative outputs $g(x) \geq 0$ are "sensitive" and we seek to certify that no input within some convex uncertainty set $X \subseteq \mathbb{R}^d$ causes the output to leave the sensitive operating regime. This asymmetric setting accurately models realistic adversarial situations. For example, an adversary may seek some imperceptible attack $x \in X = \{x' \in \mathbb{R}^d : \|x' - \overline{x}\| \leq \epsilon\}$ to cause a vehicle's image classifier to predict "no pedestrian" ($g(x) < 0$) when the nominal image $\overline{x}$ has a pedestrian in view (the sensitive regime; $g(x) \geq 0$), but not the other way around. We leave as future work the extension to vector-valued models.

Formally, the certification problem we seek to solve in this work is written

$$p^\star := \inf_{x \in X} g(x).$$

The model $g$ is robust if and only if $p^\star \geq 0$. On the other hand, if $x^\star$ solves $p^\star$, then $x^\star$ is an optimal (worst-case) attack in $X$, and it is successful if $p^\star < 0$.

The problem $p^\star$ is nonconvex due to the nonconvexity of $g$. When $g$ is a min-max affine function, a naive reformulation of $p^\star$ yields that

$$p^\star = \inf_{(x,i,t) \in X \times \mathcal{I} \times \mathbb{R}} \{t : a_{ij}^\top x + b_{ij} \leq t \text{ for all } j \in \mathcal{J}\},$$

which removes the nonconvexity in $x$ but is inefficient to solve in general due to the integer variable $i$. Alternatively, one may attempt to directly reformulate the problem into a convex one by minimizing the convex envelope of $g$. Although the resulting problem coincides with our convex reformulation $\underline{c}$ (introduced in Section III-A) on the relative interior of the direct reformulation's feasible set, it is difficult to obtain regularity conditions under which the direct reformulation holds with respect to its entire feasible set.

We propose an alternative approach to solving $p^\star$ that consists of three steps: 1) lift the problem to an optimization over probability measures, 2) leverage results and regularity conditions in distributionally robust optimization to make a finite-dimensional reduction of the problem, and 3) reformulate and solve the finite-dimensional reduction.

### A. Lifting the Problem

We lift the problem to an optimization over probability measures by noting that $g(x) = \int_X g(x') d\delta_x(x') = \mathbb{E}_{x' \sim \delta_x} g(x')$ whenever $x \in X$:

$$p^\star = \inf_{\delta_x \in \mathcal{D}(X)} \mathbb{E}_{x' \sim \delta_x} g(x').$$

With this reformulation, the optimization objective is linear in the variable $\delta_x$, but the feasible set $\mathcal{D}(X)$ is nonconvex, making the problem intractable as written. Therefore, we consider relaxing the problem to an optimization over all probability measures:

$$p' := \inf_{\mu \in \mathcal{P}(X)} \mathbb{E}_{x' \sim \mu} g(x').$$

The problem $p'$ is convex, but infinite-dimensional. We start by showing that the relaxation is exact:

**Proposition 1.** *It holds that $p' = p^\star$.*

*Proof.* Since $\mathcal{D}(X) \subseteq \mathcal{P}(X)$, it holds that $p' \leq p^\star$. Now, let $\mu \in \mathcal{P}(X)$. Then, since $p^\star \leq g(x')$ for all $x' \in X$, it holds that

$$p^\star = \int_X p^\star d\mu(x') \leq \int_X g(x') d\mu(x') = \mathbb{E}_{x' \sim \mu} g(x').$$

Since $\mu \in \mathcal{P}(X)$ is arbitrary, we conclude that $p^\star \leq \inf_{\mu \in \mathcal{P}(X)} \mathbb{E}_{x' \sim \mu} g(x') = p'$. Hence, $p' = p^\star$. □

Next, we show that solutions of the nonconvex problem $p^\star$ are generated by discrete solutions of the relaxation $p'$.

**Proposition 2.** *If $\mu^\star = \sum_{i \in \mathcal{I}} \lambda_i \delta_{x_i}$ is a discrete probability measure that solves $p'$, then $x^\star := x_i$ solves $p^\star$ for all $i \in \mathcal{I}$ such that $\lambda_i > 0$.*

*Proof.* Let $i^\star \in \arg\min_{i \in \mathcal{I}} g(x_i)$. Since $\lambda_i \geq 0$ for all $i$, $\sum_{i \in \mathcal{I}} \lambda_i = 1$, and $g(x_{i^\star}) \leq g(x_i)$ for all $i$, it holds that

$$\begin{aligned} p^\star \leq g(x_{i^\star}) &= \sum_{i \in \mathcal{I}} \lambda_i g(x_{i^\star}) \\ &\leq \sum_{i \in \mathcal{I}} \lambda_i g(x_i) = \mathbb{E}_{x' \sim \mu^\star} g(x') = p', \end{aligned}$$

so $x_{i^\star}$ solves $p^\star$ by Proposition 1. If $x_{i'}$ does not solve $p^\star$ for some $i' \in \mathcal{I}$ such that $\lambda_{i'} > 0$, then $p^\star = g(x_{i^\star}) < g(x_{i'})$, implying that $\sum_{i \in \mathcal{I}} \lambda_i g(x_{i^\star}) < \sum_{i \in \mathcal{I}} \lambda_i g(x_i)$ and hence that $p^\star < p'$, which contradicts Proposition 1. □

The above results show that we may solve the problem $p^\star$ of interest by solving $p'$ for a discrete optimal distribution. The remainder of this section is dedicated to this approach.

### B. Finite-Dimensional Reduction

To make our finite-dimensional reduction, we recall the definitions of conjugate and perspective functions.

**Definition 2.** The *conjugate* of a function $f \colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is the function $f^* \colon \mathbb{R}^d \to \overline{\mathbb{R}}$ defined by

$$f^*(y) = \sup_{x \in \operatorname{dom}(f)} (y^\top x - f(x)).$$

We write $f^{**}$ to denote the biconjugate $(f^*)^*$.

**Definition 3.** The *perspective* of a proper, closed, and convex function $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is the function $\mathscr{P}_f\colon \mathbb{R}^d \times \mathbb{R}_+ \to \overline{\mathbb{R}}$ defined by

$$\mathscr{P}_f(x,t) = \begin{cases} tf(x/t) & \text{if } t>0, \\ \sup_{y\in\operatorname{dom}(f^*)} y^\top x & \text{if } t=0. \end{cases}$$

Recall that the perspective $\mathscr{P}_f$ of a convex function $f$ is also convex, and that the conjugate $f^*$ is convex even when $f$ is nonconvex [39].

Throughout the remainder of the paper, we fix $g$ and $X$ to be min-max affine and convex, respectively, via the following structural assumptions:

**Assumption 1.** It holds that $g \in \mathcal{G}$, taking the form $g(x) = \min_{i\in\mathcal{I}} g_i(x)$ with $g_i(x) = \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij})$.

**Assumption 2.** The set $X$ takes the form $X = \{x \in \mathbb{R}^d : c_k(x) \leq 0,\ k\in\mathcal{K}\}$ with $c_k\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ a proper, closed, and convex function for all $k \in \mathcal{K}$.

We now make the reduction by introducing two finite-dimensional convex optimization problems:

$$\begin{aligned} \underline{c} := \underset{\substack{\lambda_i,\eta_i\in\mathbb{R}\\ x_i\in\mathbb{R}^d}}{\text{minimize}} \quad & \sum_{i\in\mathcal{I}} \eta_i \\ \text{subject to} \quad & \mathscr{P}_{c_k}(x_i,\lambda_i) \leq 0,\ i\in\mathcal{I},\ k\in\mathcal{K}, \\ & \mathscr{P}_{g_i}(x_i,\lambda_i) \leq \eta_i,\ i\in\mathcal{I}, \\ & \sum_{i\in\mathcal{I}} \lambda_i = 1,\ \lambda \geq 0. \end{aligned}$$

$$\begin{aligned} \overline{c} := \underset{\substack{\alpha,\beta_{ik}\in\mathbb{R}\\ y_i,z_{ik}\in\mathbb{R}^d}}{\text{maximize}} \quad & -\alpha \\ \text{subject to} \quad & g_i^*(y_i) + \sum_{k\in\mathcal{K}} \mathscr{P}_{c_k^*}(z_{ik},\beta_{ik}) \leq \alpha,\ i\in\mathcal{I}, \\ & y_i + \sum_{k\in\mathcal{K}} z_{ik} = 0,\ i\in\mathcal{I}, \\ & \beta_{ik} \geq 0,\ i\in\mathcal{I},\ k\in\mathcal{K}. \end{aligned}$$

Intuitively, $\underline{c}$ is minimizing a sort of "average" of the components $g_i$ at a finite number of points $x_i$ with weights given by the probability vector $\lambda$, and $\overline{c}$ is its dual. We now leverage recent results in distributionally robust optimization to show that the finite reductions $\underline{c}, \overline{c}$ allow us to solve the infinite-dimensional problem $p'$ under mild assumptions.

**Definition 4.** Let $f_0, f_1, \dots, f_m$ and $h_1, \dots, h_n$ be extended real-valued functions defined on $\mathbb{R}^d$. The optimization problem $p = \inf\{f_0(x) : f_1(x) \leq 0, \dots, f_m(x) \leq 0,\ h_1(x) = 0, \dots, h_n(x) = 0,\ x \in \mathbb{R}^d\}$ *admits a Slater point* if there exists $x \in \bigcap_{i=0}^m \operatorname{ri}(\operatorname{dom}(f_i)) \cap \bigcap_{j=1}^n \operatorname{ri}(\operatorname{dom}(h_j))$ such that $f_i(x) \leq 0$ and $h_j(x) = 0$ for all $i$ and all $j$, and such that $f_i(x) < 0$ for all $i \neq 0$ such that $f_i$ is nonlinear.

**Assumption 3.** The set $X$ is bounded and the optimization problem $\overline{c}$ admits a Slater point.

The above boundedness assumption on $X$ is standard in the adversarial robustness literature. The Slater condition may be verified by simply solving $\overline{c}$ with a small number $\epsilon > 0$ added to all of the nonlinear inequality constraints; replace $f_i(x) \leq 0$ with $f_i(x) + \epsilon \leq 0$ for all nonlinear constraint functions $f_i$.

**Theorem 2.** *If Assumption 3 holds, then $\underline{c} = p' = \overline{c}$, and the discrete probability distribution $\sum_{i\in\mathcal{I}:\lambda_i^\star\neq 0} \lambda_i^\star \delta_{x_i^\star/\lambda_i^\star}$ solves $p'$ for all solutions $(\eta^\star, \lambda^\star, x^\star)$ to $\underline{c}$.*

*Proof.* Since $X$ is defined by a finite intersection of 0-sublevel sets of proper, closed, and convex functions (Assumption 2), and since every $g_i\colon x \mapsto \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij})$ is a proper, closed, and convex function, the result follows from [40, Theorem 12(ii)]. □

Theorem 2 together with our Propositions 1 and 2 show that we are able to *exactly* compute an optimal attack solving the nonconvex problem $p^\star$ by solving the convex optimizations $\underline{c}, \overline{c}$.

### C. Reformulating and Solving the Finite Reduction

In order to solve $\underline{c}, \overline{c}$, we must derive the appropriate conjugates and perspectives. In this subsection, we do so for the common cases where $X$ is defined in terms of norm balls or polyhedra. We will also see that computing the conjugate $g_i^*$ is highly nontrivial, and as a result we turn to tractably reformulating the constraint involving $g_i^*$ using duality theory.

**Proposition 3.** *The perspective of $g_i\colon x \mapsto \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij})$ is given by $\mathscr{P}_{g_i}(x,t) = \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij}t)$ for all $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$.*

*Proof.* Let $x \in \mathbb{R}^d$. If $t > 0$, then

$$\begin{aligned} \mathscr{P}_{g_i}(x,t) &= tg_i(x/t) \\ &= t\max_{j\in\mathcal{J}}(a_{ij}^\top x/t + b_{ij}) = \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij}t). \end{aligned}$$

If $t = 0$, then

$$\begin{aligned} \mathscr{P}_{g_i}(x,t) &= \liminf_{(x',t')\to(x,0)} \mathscr{P}_{g_i}(x',t') \\ &= \liminf_{(x',t')\to(x,0)} \max_{j\in\mathcal{J}}(a_{ij}^\top x' + b_{ij}t') \\ &= \max_{j\in\mathcal{J}} a_{ij}^\top x = \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij}t), \end{aligned}$$

where the first equality comes from Theorem 13.3 and Corollary 8.5.2 in [27] and the third equality comes from the continuity of $(x,t) \mapsto \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij}t)$. □

**Proposition 4.** *The perspective of $c_k\colon x \mapsto \|x - \overline{x}\| - \epsilon$ is given by $\mathscr{P}_{c_k}(x,t) = \|x - t\overline{x}\| - \epsilon t$ for all $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$.*

*Proof.* Following the same reasoning as in the proof of Proposition 3, we find that $\mathscr{P}_{c_k}(x,t) = t(\|x/t - \overline{x}\| - \epsilon) = \|x - t\overline{x}\| - \epsilon t$ for $t > 0$ and $\mathscr{P}_{c_k}(x,t) = \liminf_{(x',t')\to(x,0)}(\|x' - t'\overline{x}\| - \epsilon t') = \|x\| = \|x - t\overline{x}\| - \epsilon t$ for $t = 0$. □

**Proposition 5.** *The conjugate of $c_k\colon x \mapsto \|x-\overline{x}\| - \epsilon$ is given for all $z \in \mathbb{R}^d$ by*

$$c_k^*(z) = \begin{cases} z^\top \overline{x} + \epsilon & \textit{if } \|z\|_* \le 1, \\ \infty & \textit{if } \|z\|_* > 1. \end{cases}$$

*Proof.* Let $z \in \mathbb{R}^d$ be such that $\|z\|_* \le 1$. Then

$$\sup_{x \in \mathbb{R}^d : x \neq \overline{x}} \frac{z^\top (x - \overline{x})}{\|x - \overline{x}\|} = \sup_{x' \in \mathbb{R}^d : \|x'\| \le 1} z^\top x' = \|z\|_* \le 1,$$

so $z^\top (x-\overline{x}) - \|x-\overline{x}\| \le 0$ for all $x \neq \overline{x}$. Also, $z^\top (x - \overline{x}) - \|x - \overline{x}\| = 0$ for $x = \overline{x}$, and therefore $\sup_{x\in\mathbb{R}^d}(z^\top(x - \overline{x}) - \|x-\overline{x}\|) = 0$, indicating that

$$c_k^*(z) = \sup_{x \in \mathbb{R}^d} \left( z^\top (x - \overline{x}) - \|x - \overline{x}\| \right) + z^\top \overline{x} + \epsilon = z^\top \overline{x} + \epsilon.$$

On the other hand, let $z \in \mathbb{R}^d$ be such that $\|z\|_* > 1$. Then there exists $x' \in \mathbb{R}^d \setminus \{0\}$ such that $\frac{z^\top x'}{\|x'\|} > 1$, implying that $z^\top x' - \|x'\| > 0$, and hence

$$\begin{aligned} c_k^*(z) &\ge z^\top (\overline{x} + \alpha x') - \|\alpha x'\| + \epsilon \\ &= \alpha (z^\top x' - \|x'\|) + z^\top \overline{x} + \epsilon \to \infty \end{aligned}$$

as $\alpha \to \infty$. Thus, $c_k^*(z) = \infty$. □

**Proposition 6.** *The perspective of the conjugate of $c_k\colon x \mapsto \|x - \overline{x}\| - \epsilon$ is given for all $(z,t) \in \mathbb{R}^d \times \mathbb{R}_+$ by*

$$\mathscr{P}_{c_k^*}(z,t) = \begin{cases} z^\top \overline{x} + \epsilon t & \textit{if } \|z\|_* \le t, \\ \infty & \textit{if } \|z\|_* > t. \end{cases}$$

*Proof.* Let $t > 0$. If $z \in \mathbb{R}^d$ is such that $\|z\|_* \le t$, then $\|z/t\|_* \le 1$, so $\mathscr{P}_{c_k^*}(z,t) = t c_k^*(z/t) = t((z/t)^\top \overline{x} + \epsilon) = z^\top \overline{x} + \epsilon t$. If $\|z\|_* > t$, then $\|z/t\|_* > 1$, so $\mathscr{P}_{c_k^*}(z,t) = t c_k^*(z/t) = \infty$.

On the other hand, let $t = 0$. Then

$$\mathscr{P}_{c_k^*}(z,t) = \sup_{x \in \operatorname{dom}(c_k^{**})} z^\top x = \sup_{x \in \mathbb{R}^d} z^\top x = \begin{cases} 0 & \text{if } z = 0, \\ \infty & \text{if } z \neq 0, \end{cases}$$

since $c_k^{**} = c_k$ which has domain $\mathbb{R}^d$, as $c_k$ is proper, closed, and convex [27, Theorem 12.2]. Since, when $t = 0$, the condition $z = 0$ is equivalent to $\|z\|_* \le t$ and the condition $z \neq 0$ is equivalent to $\|z\|_* > t$, the proof is complete. □

We also provide the conjugates and perspectives for polyhedral $X$:

**Proposition 7.** *Let $c_k\colon x \mapsto \psi_k^\top x + \omega_k$ for some $\psi_k \in \mathbb{R}^d$ and some $\omega_k \in \mathbb{R}$. Then the following all hold:*

1) $\mathscr{P}_{c_k}(x,t) = \psi_k^\top x + \omega_k t,$
2) $c_k^*(z) = \begin{cases} -\omega_k & \textit{if } z = \psi_k, \\ \infty & \textit{if } z \neq \psi_k, \end{cases}$
3) *and* $\mathscr{P}_{c_k^*}(z,t) = \begin{cases} -\omega_k t & \textit{if } z = t\psi_k, \\ \infty & \textit{if } z \neq t\psi_k. \end{cases}$

The proof of Proposition 7 follows from a straightforward application of the definitions of conjugate and perspective, and is hence omitted for brevity.

The conjugate $g_i^*$ is all that remains to compute. However, although computing $g_i^*$ in closed form for the univariate ($d = 1$) function $g_i\colon x \mapsto \max_{j\in\mathcal{J}}(a_{ij}x + b_{ij})$ can be straightforward, generalizing the formula to higher-dimensional settings is nontrivial. In theory, it is possible to express $g_i^*$ for $d > 1$ in closed form via [27, Theorem 19.2]. However, this requires solving a vertex enumeration problem, i.e., determining finite sets $V, R \subseteq \mathbb{R}^d \times \mathbb{R}$ such that the polyhedron $\operatorname{epi}(g_i^*) \coloneqq \{(x,t) \in \mathbb{R}^d \times \mathbb{R} : a_{ij}^\top x + b_{ij} \le t \text{ for all } j \in \mathcal{J}\}$ equals $\operatorname{conv}(P) + \operatorname{cone}(R)$. The vertex enumeration problem is NP-hard in general [41]. See the Minkowski-Weyl theorem [27, Theorem 19.1] for the theory on such representations of polyhedra. In Theorem 3 that follows, we instead take a duality-based robust optimization approach to tractably deal with the conjugate $g_i^*$ in a direct manner.

**Lemma 1.** *It holds that $\operatorname{dom}(g_i^*) = \operatorname{conv}\{a_{ij} : j \in \mathcal{J}\}$.*

*Proof.* Let $y \in \operatorname{conv}\{a_{ij} : j \in \mathcal{J}\}$. Then $y = \sum_{j\in\mathcal{J}} \theta_j a_{ij}$ for some $\theta \in \mathbb{R}^n$ such that $\theta \ge 0$ and $\sum_{j\in\mathcal{J}} \theta_j = 1$. Hence, for all $x \in \mathbb{R}^d$, we find that

$$\begin{aligned} y^\top x - \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij}) &= \sum_{j\in\mathcal{J}} \theta_j a_{ij}^\top x - \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij}) \\ &= \sum_{j\in\mathcal{J}} \theta_j (a_{ij}^\top x + b_{ij}) \\ &\quad - \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij}) - \sum_{j\in\mathcal{J}} \theta_j b_{ij} \\ &\le \sum_{j\in\mathcal{J}} \theta_j \max_{l\in\mathcal{J}}(a_{il}^\top x + b_{il}) \\ &\quad - \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij}) - \sum_{j\in\mathcal{J}} \theta_j b_{ij} \\ &= -\sum_{j\in\mathcal{J}} \theta_j b_{ij}, \end{aligned}$$

and thus $g_i^*(y) \le -\sum_{j\in\mathcal{J}} \theta_j b_{ij} < \infty$, so $y \in \operatorname{dom}(g_i^*)$.

On the other hand, let $y \in \operatorname{dom}(g_i^*)$, so that $g_i^*(y) < \infty$. An epigraphic reformulation of $g_i^*(y)$ yields that $\infty > g_i^*(y) = \sup_{x\in\mathbb{R}^d}(y^\top x - \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij})) = \sup_{(x,t)\in\mathbb{R}^d\times\mathbb{R}}\{y^\top x - t : a_{ij}^\top x + b_{ij} \le t \text{ for all } j \in \mathcal{J}\}$. This reformulation is a linear program with a finite optimal value, and hence by [42, Proposition 3.1.3], the reformulation is attained by some $(x,t) \in \mathbb{R}^d \times \mathbb{R}$, and since it must be the case that $t = a_{ij}^\top x + b_{ij}$ for some $j \in \mathcal{J}$ at this point $(x,t)$, we conclude that this $x$ solves the supremum defining $g_i^*(y)$ in its original form (i.e., pre-epigraphic reformulation). Therefore, by the first-order optimality condition for unconstrained convex optimization [27, Theorem 23.2], it holds that $0 \in \partial h_i(x)$, where $h_i\colon \mathbb{R}^d \to \mathbb{R}$ is the convex function defined by $h_i(x) = \max_{j\in\mathcal{J}}(a_{ij}^\top x + b_{ij}) - y^\top x$. Using the rules for subdifferentials of pointwise maxima and sums of proper convex functions [42, Proposition B.22],[27, Theorem 23.8], we have that $\partial h_i(x) = \operatorname{conv}\left(\bigcup_{j\in\mathcal{A}(x)}\{a_{ij}\}\right) + \{-y\}$, where $\mathcal{A}(x)$ denotes the set of active indices at $x$: $\mathcal{A}(x) = \{j \in \mathcal{J} : a_{ij}^\top x + b_{ij} = \max_{l\in\mathcal{J}}(a_{il}^\top x + b_{il})\}$. Since $0 \in \partial h_i(x)$, this yields that $y \in \operatorname{conv}\left(\bigcup_{j\in\mathcal{A}(x)}\{a_{ij}\}\right) \subseteq \operatorname{conv}\{a_{ij} : j \in \mathcal{J}\}$. This completes the proof. □

**Assumption 4.** The functions $g_i$ are nonredundant in the sense that for all $j \in \mathcal{J}$ there exists $x \in \mathbb{R}^d$ such that $g_i(x) = a_{ij}^\top x + b_{ij}$.

It is easy to see that nonredundancy of $g_i$ is efficiently verified by solving the linear (feasibility) programs $\inf\{0 : (a_{il} - a_{ij})^\top x + (b_{il} - b_{ij}) \leq 0 \text{ for all } l \in \mathcal{J},\ x \in \mathbb{R}^d\}$ for all $j \in \mathcal{J}$. Removing the affine components of $g_i$ with infeasible programs ensures that Assumption 4 holds and does not change the model's predictions.

**Theorem 3.** *Suppose that Assumption 4 holds, and let $h \colon \Gamma \to \mathbb{R}$ be an arbitrary real-valued function defined on some nonempty set $\Gamma$. Then, for all $y \in \mathbb{R}^d$ and all $\gamma \in \Gamma$, it holds that $g_i^*(y) \leq h(\gamma)$ if and only if, for all $j \in \mathcal{J}$, there exists $\nu_{ij} \in \mathbb{R}^n$ such that the following all hold:*

1) *$y = \sum_{j \in \mathcal{J}} \theta_j a_{ij}$ for some $\theta \in \mathbb{R}^n$ such that $\theta \geq 0$ and $\sum_{j \in \mathcal{J}} \theta_j = 1$,*
2) *$\nu_{ij} \geq 0$,*
3) *$y - a_{ij} + \sum_{l \in \mathcal{J}} (\nu_{ij})_l (a_{ij} - a_{il}) = 0$,*
4) *and $-b_{ij} + \sum_{l \in \mathcal{J}} (\nu_{ij})_l (b_{ij} - b_{il}) \leq h(\gamma)$.*

*Proof.* Let $y \in \mathbb{R}^d$ and $\gamma \in \Gamma$. If $y \neq \sum_{j \in \mathcal{J}} \theta_j a_{ij}$ for all $\theta \in \mathbb{R}^n$ such that $\theta \geq 0$ and $\sum_{j \in \mathcal{J}} \theta_j = 1$, then $y \notin \operatorname{conv}\{a_{ij} : j \in \mathcal{J}\}$ and hence $y \notin \operatorname{dom}(g_i^*)$ by Lemma 1. In this case, $g_i^*(y) = \infty > h(\gamma)$ since $h$ is real-valued. Therefore, the first condition enumerated in the theorem is necessary for $g_i^*(y) \leq h(\gamma)$.

Going forward, assume that $y = \sum_{j \in \mathcal{J}} \theta_j a_{ij}$ for some $\theta \in \mathbb{R}^n$ such that $\theta \geq 0$ and $\sum_{j \in \mathcal{J}} \theta_j = 1$. Hence, $g_i^*(y) < \infty$. Breaking up the conjugate's supremum into $n$ suprema over the affine components of $g_i$ yields

$$\begin{aligned} g_i^*(y) &= \sup_{x \in \mathbb{R}^d} (y^\top x - \max_{j \in \mathcal{J}} (a_{ij}^\top x + b_{ij})) \\ &= \max_{j \in \mathcal{J}} \sup_{x \in \mathbb{R}^d} \{(y - a_{ij})^\top x - b_{ij} : \\ &\qquad (a_{il} - a_{ij})^\top x + (b_{il} - b_{ij}) \leq 0 \text{ for all } l \in \mathcal{J}\}. \end{aligned}$$

Denote the inner suprema by $\mathfrak{p}_{ij} \coloneqq \sup_{x \in \mathbb{R}^d} \{(y - a_{ij})^\top x - b_{ij} : (a_{il} - a_{ij})^\top x + (b_{il} - b_{ij}) \leq 0 \text{ for all } l \in \mathcal{J}\}$. Since, by Assumption 4, for all $j \in \mathcal{J}$ there exists $x \in \mathbb{R}^d$ such that $\max_{l \in \mathcal{J}} (a_{il}^\top x + b_{il}) = g_i(x) = a_{ij}^\top x + b_{ij}$, it holds that $\{x \in \mathbb{R}^d : a_{ij}^\top x + b_{ij} \geq a_{il}^\top x + b_{il} \text{ for all } l \in \mathcal{J}\} \neq \emptyset$ for all $j \in \mathcal{J}$, implying that every $\mathfrak{p}_{ij}$ is feasible, i.e., $\mathfrak{p}_{ij} > -\infty$. Furthermore, since $g_i^*(y) < \infty$, it must be the case that $\mathfrak{p}_{ij} < \infty$ for all $j \in \mathcal{J}$. Thus, every optimal value $\mathfrak{p}_{ij}$ is finite. Therefore, by [42, Proposition 3.1.3], every $\mathfrak{p}_{ij}$ is attained, and therefore by [42, Proposition 4.4.2] strong duality holds between $\mathfrak{p}_{ij}$ and its dual problem, which we denote by $\mathfrak{d}_{ij}$, and it also holds that $\mathfrak{d}_{ij}$ is attained. A routine derivation via Lagrangian duality therefore yields that

$$\begin{aligned} \mathfrak{p}_{ij} &= \mathfrak{d}_{ij} \\ &= \inf_{\nu_{ij} \in \mathbb{R}^n} \Bigg\{ \sum_{l \in \mathcal{J}} (\nu_{ij})_l (b_{ij} - b_{il}) - b_{ij} : \\ &\qquad y - a_{ij} + \sum_{l \in \mathcal{J}} (\nu_{ij})_l (a_{ij} - a_{il}) = 0,\ \nu_{ij} \geq 0 \Bigg\}. \end{aligned}$$

Hence, $g_i^*(y) \leq h(\gamma)$ if and only if $\max_{j \in \mathcal{J}} \mathfrak{p}_{ij} \leq h(\gamma)$ if and only if $\mathfrak{p}_{ij} \leq h(\gamma)$ for all $j \in \mathcal{J}$. Thus, since $\mathfrak{d}_{ij}$ is attained, it holds that $g_i^*(y) \leq h(\gamma)$ if and only if, for all $j \in \mathcal{J}$, there exists $\nu_{ij} \in \mathbb{R}^n$ such that $\nu_{ij} \geq 0$, $y - a_{ij} + \sum_{l \in \mathcal{J}} (\nu_{ij})_l (a_{ij} - a_{il}) = 0$, and $-b_{ij} + \sum_{l \in \mathcal{J}} (\nu_{ij})_l (b_{ij} - b_{il}) \leq h(\gamma)$. This completes the proof. □

With the above conjugate and perspective derivations, our reformulations of $\underline{c}, \overline{c}$ are complete; they may now be directly solved using off-the-shelf convex optimization solvers.

*Remark* 1. Our developments can be generalized, so long as one can compute the appropriate conjugates and perspectives. In particular, the mathematical machinery yielding a discrete distribution solution to $p'$ from a solution to an associated finite-dimensional convex optimization problem may be applied to general convex functions $g_i$ and other (non-norm-based and non-polyhedral) convex attack sets $X$ [40]. In fact, moment constraints on $\mu \in \mathcal{P}(X)$ may even be added to the semi-infinite program $p'$, which may allow for modeling alternative "distributional attacks" beyond the standard "Dirac attack" at a single point considered here.

## IV. Experiments

In this section, we illustrate the utility of our method in both robust control and image classification settings.[3]

### A. Robust Control Certification

We take an illustrative robust control example adapted from the well-known autonomous vehicle collision avoidance problem [43], [44]. Consider two planar vehicles approaching an intersection located at the origin $(0, 0) \in \mathbb{R}^2$. One vehicle travels east with state $\mathbf{x}(t) = (x(t), \dot{x}(t)) \in \mathbb{R}^2$ at time $t$. The other vehicle, which we control and hence term the "ego vehicle," travels north with state $\mathbf{y}(t) = (y(t), \dot{y}(t))$. The eastbound uncontrolled vehicle has a fixed velocity ($\ddot{x}(t) = 0$ for all $t$). The full state $(x(t), \dot{x}(t), y(t), \dot{y}(t))$ is randomly initialized at $t = 0$ within $[-3, -2] \times [1/2, 5/2] \times [-3, -2] \times [0, 2]$. The vehicles are each $1$ unit long and $1/2$ unit wide, matching the width of the road. Thus, a vehicle is considered to be in the intersection if the absolute value of its position is less than $3/4$. If the vehicles collide, the simulation is stopped. We simulate standard double integrator dynamics with a time step $\Delta t = 0.05$ for 100 steps.

We control the northbound vehicle using a learned policy $u(t) = -\pi_\theta(\mathbf{x}(t), \mathbf{y}(t))$ that enters the dynamics as $\ddot{y}(t) = \Pi_{[-1,1]}(u(t))$, where $\pi_\theta \colon \mathbb{R}^4 \to \mathbb{R}$ is a min-max affine function with $m = n = 10$ and $\Pi_{[-1,1]}$ is the natural projection mapping of $\mathbb{R}$ onto $[-1, 1]$. Our robustness certificates apply for all training schemes, e.g., reinforcement learning and imitation learning. We train $\pi_\theta$ using imitation learning on 500 trajectories generated by a hand-programmed expert policy $\pi^\star$. We use the mean squared error loss function and train for 20 epochs using the Adam optimizer at a learning rate of 0.01. The expert policy $\pi^\star$ is designed to stop the ego vehicle $\delta = 0.1$ units before the intersection with a

[3] All experiments are conducted on a Ubuntu 22.04 instance with an Intel i7-9700K CPU and NVIDIA RTX A6000 GPU.

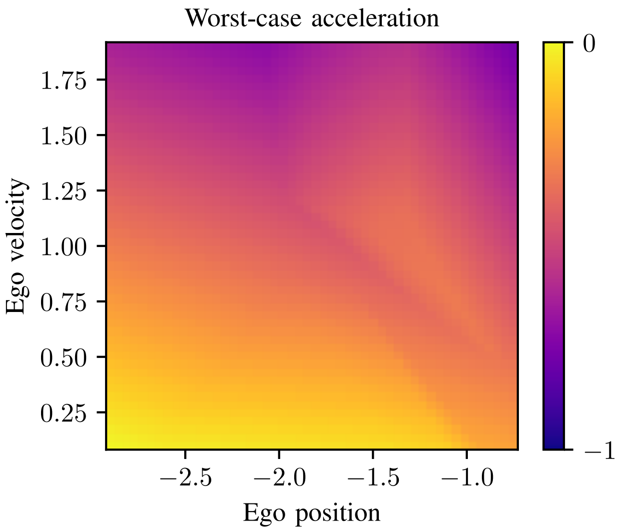

Fig. 1. Largest possible acceleration over all uncontrolled vehicle states for particular values of the ego vehicle state. The output is always negative, ensuring some level of braking.

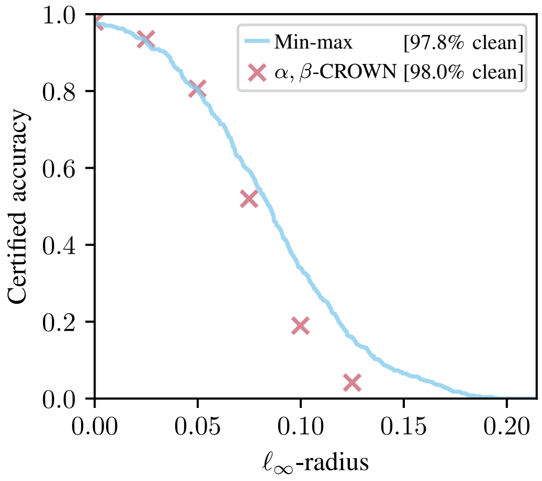

Fig. 2. Certified accuracies of our min-max representation and of $\alpha, \beta$-CROWN on the MNIST 3-versus-8 dataset.

constant acceleration, then apply no acceleration until the tail of the uncontrolled vehicle is $\delta$ units past the intersection, and then accelerate with the maximum input of 1.

We now consider certifying the safety of our control system. Our goal is to guarantee that the ego vehicle always brakes when the uncontrolled vehicle is approaching or inside the intersection. This enforcement of braking corresponds to ensuring that the largest acceleration signal $u(t)$ is less than zero, which amounts to minimizing the output of $\pi_\theta$ over the set of states for which we desire braking. This is formalized by requiring braking for all states in the set

$$X = [-3+\delta, \tfrac{3}{4}] \times [\tfrac{1}{2}+\delta, \tfrac{5}{2}-\delta] \times [-3+\delta, -\tfrac{3}{4}] \times [\delta, 2-\delta],$$

which consists of states where the uncontrolled vehicle is approaching or in the intersection and the ego vehicle is approaching the intersection. The small positive constant $\delta = 0.1$ accounts for boundary states where expert trajectories may not have been sampled.

Utilizing our robustness certificates from Section III, we verify that indeed $u(t) = -\pi_\theta\left(\mathbf{x}(t), \mathbf{y}(t)\right) < 0$ for all states $(\mathbf{x}(t), \mathbf{y}(t)) \in X$. For visual purposes, we also consider fixing a particular $\mathbf{y}(t)$ and computing the largest possible acceleration $u(t)$ amongst all uncontrolled vehicle states $\mathbf{x}(t)$ captured by $X$. The solutions to this problem over a range of $\mathbf{y}(t)$ are plotted in Figure 1. As expected, for all $\mathbf{y}(t)$, the ego vehicle is braking. As the ego vehicle approaches the intersection (large $y(t)$) or becomes faster (large $\dot{y}(t)$), we certify that the controller brakes more heavily.

### B. Image Classification

We demonstrate the tightness and efficiency of our method on an image classification example adapted from [38]. The task is to distinguish between two visually similar MNIST classes: the digits 3 and 8 [45]. As we consider the asymmetric setting, we aim to certify predictions for one particular class, which we take to be the class of 3's, while maintaining high clean accuracy for both classes. We consider the attack set $X = \{x \in \mathbb{R}^d : \|x - \overline{x}\|_\infty \leq \epsilon\}$ over a range of radii $\epsilon > 0$ around test images $\overline{x}$. In this setting, certificates ensure that pixelwise adversarial alterations of an image $\overline{x}$ of a 3 cannot fool the classifier into predicting an 8.

We compare two approaches: 1) directly learning our min-max representation with $n = m = 15$ and certifying via our convex optimization-based certificates, and 2) learning a standard composition-based ReLU model and certifying via the state-of-the-art verifier $\alpha, \beta$-CROWN [22]. Since $\alpha, \beta$-CROWN's worst-case runtime scales exponentially with model size, we instantiate the standard ReLU model with one hidden layer and 100 hidden units, which is the smallest hidden layer size that yields comparable clean accuracy to our min-max representation. We use adversarial training (see [9]) with $\ell_\infty$-attacks starting at a radius of 0.001 and linearly interpolate to a radius of $\epsilon_{\text{train}}$ over the first 20 epochs, where $\epsilon_{\text{train}} = 0.05$ for our model and $\epsilon_{\text{train}} = 0.3$ for the standard ReLU model. Both models are trained using the Adam optimizer with a learning rate of 0.001 for 60 epochs.

Figure 2 compares the certified accuracy (averaged over the test inputs) of our method against that of $\alpha, \beta$-CROWN. As certifying at a particular $\epsilon$ using our method is fast, for each test input, the largest certifiable $\ell_\infty$-radius is found using binary search in order to yield a smooth certified accuracy curve. On the other hand, due to the expensive runtime of $\alpha, \beta$-CROWN, we only certify at the select radii shown. Our min-max representation exceeds the state-of-the-art baseline certified radii at far faster runtimes: certifying a single input-radius pair $(\overline{x}, \epsilon)$ takes on average 3.67 seconds with $\alpha, \beta$-CROWN versus only 0.48 seconds with our method. We note that our runtime comparisons are solely based off of models with equivalent clean accuracy. Due to space constraints, we leave more thorough analyses of relative expressivity and computational complexity for future work.

## V. Conclusions

In this work, we *exactly* solve the nonconvex robustness certification problem over convex attack sets for min-max representations of ReLU neural networks by developing a tractable convex reformulation. An interesting line of future work may include developing more efficient min-max representations or estimations for arbitrary ReLU neural networks, so that the advantageous optimization properties derived in this paper may be easily applied. Other interest lies in comparing the number of affine regions of a general min-max affine function versus that of a general ReLU neural network.

## References

[1] S. Chen, S. A. Billings, and P. Grant, "Non-linear system identification using neural networks," *International Journal of Control*, 1990.

[2] S. S. Ge, C. C. Hang, T. H. Lee, and T. Zhang, *Stable Adaptive Neural Network Control*, 2013.

[3] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, 2016.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[5] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[6] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, "Perceptual-sensitive GAN for generating adversarial patches," in *AAAI Conference on Artificial Intelligence*, 2019.

[7] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[8] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient SMT solver for verifying deep neural networks," in *Computer Aided Verification: 29th International Conference*, 2017.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[10] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, 2017.

[11] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International Conference on Machine Learning*, 2018.

[12] H. Zhang, T.-W. Weng, P.-Y. Chen, C.-J. Hsieh, and L. Daniel, "Efficient neural network robustness certification with general activation functions," in *Advances in Neural Information Processing Systems*, 2018.

[13] A. Raghunathan, J. Steinhardt, and P. S. Liang, "Semidefinite relaxations for certifying robustness to adversarial examples," in *Advances in Neural Information Processing Systems*, 2018.

[14] M. Fazlyab, M. Morari, and G. J. Pappas, "Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming," *IEEE Transactions on Automatic Control*, 2020.

[15] B. G. Anderson, Z. Ma, J. Li, and S. Sojoudi, "Towards optimal branching of linear and semidefinite relaxations for neural network robustness certification," *arXiv preprint arXiv:2101.09306*, 2023.

[16] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. Pappas, "Efficient and accurate estimation of Lipschitz constants for deep neural networks," in *Advances in Neural Information Processing Systems*, 2019.

[17] L. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, L. Daniel, D. Boning, and I. Dhillon, "Towards fast computation of certified robustness for ReLU networks," in *International Conference on Machine Learning*, 2018.

[18] A. Virmaux and K. Scaman, "Lipschitz regularity of deep neural networks: Analysis and efficient estimation," in *Advances in Neural Information Processing Systems*, 2018.

[19] M. Jordan and A. G. Dimakis, "Exactly computing the local Lipschitz constant of ReLU networks," in *Advances in Neural Information Processing Systems*, 2020.

[20] V. Tjeng, K. Xiao, and R. Tedrake, "Evaluating robustness of neural networks with mixed integer programming," in *International Conference on Learning Representations*, 2019.

[21] B. G. Anderson, Z. Ma, J. Li, and S. Sojoudi, "Tightened convex relaxations for neural network robustness certification," in *IEEE Conference on Decision and Control*, 2020.

[22] S. Wang, H. Zhang, K. Xu, X. Lin, S. Jana, C.-J. Hsieh, and J. Z. Kolter, "Beta-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification," in *Advances in Neural Information Processing Systems*, 2021.

[23] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," in *Advances in Neural Information Processing Systems*, 2014.

[24] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, "Understanding deep neural networks with rectified linear units," in *International Conference on Learning Representations*, 2018.

[25] L. Zhang, G. Naitzat, and L.-H. Lim, "Tropical geometry of deep neural networks," in *International Conference on Machine Learning*, 2018.

[26] P. Awasthi, A. Dutta, and A. Vijayaraghavan, "On robustness to adversarial examples and polynomial optimization," in *Advances in Neural Information Processing Systems*, 2019.

[27] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[28] J.-L. Marichal, "Weighted lattice polynomials," *Discrete Mathematics*, 2009.

[29] S. Velasco-Forero and J. Angulo, "MorphoActivation: Generalizing ReLU activation function by mathematical morphology," in *Discrete Geometry and Mathematical Morphology*, 2022.

[30] A. M. Bagirov, "Max-min separability," *Optimization Methods and Software*, 2005.

[31] A. M. Bagirov and J. Ugon, "Supervised data classification via max-min separability," *Continuous Optimization: Current Trends and Modern Applications*, 2005.

[32] B. Rister and D. L. Rubin, "Piecewise convexity of artificial neural networks," *Neural Networks*, 2017.

[33] S. Chen, A. R. Klivans, and R. Meka, "Learning deep ReLU networks is fixed-parameter tractable," *arXiv preprint arXiv:2009.13512*, 2020.

[34] S. Wang and K. S. Narendra, "Nonlinear system identification with lattice piecewise-linear functions," in *American Control Conference*, 2002.

[35] C. D. Aliprantis, D. Harris, and R. Tourky, "Riesz estimators," *Journal of Econometrics*, 2007.

[36] S. Ovchinnikov, "Max-min representation of piecewise linear functions," *Contributions to Algebra and Geometry*, 2002.

[37] A. Pinkus, "Approximation theory of the MLP model in neural networks," *Acta Numerica*, 1999.

[38] S. Pfrommer, B. G. Anderson, J. Piet, and S. Sojoudi, "Asymmetric certified robustness via feature-convex neural networks," *arXiv preprint arXiv:2302.01961*, 2023.

[39] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[40] J. Zhen, D. Kuhn, and W. Wiesemann, "Mathematical foundations of robust and distributionally robust optimization," *arXiv preprint arXiv:2105.00760*, 2021.

[41] L. Khachiyan, E. Boros, K. Borys, V. Gurvich, and K. Elbassioni, "Generating all vertices of a polyhedron is hard," *Discrete & Computational Geometry*, 2009.

[42] D. P. Bertsekas, *Nonlinear Programming*, 3rd ed. Athena Scientific, 2016.

[43] K. Ren, H. Ahn, and M. Kamgarpour, "Chance-constrained trajectory planning with multimodal environmental uncertainty," *IEEE Control Systems Letters*, 2022.

[44] A. Wang, A. Jasour, and B. C. Williams, "Non-Gaussian chance-constrained trajectory planning for autonomous vehicles under agent uncertainty," *IEEE Robotics and Automation Letters*, 2020.

[45] Y. LeCun, "The MNIST database of handwritten digits," `http://yann.lecun.com/exdb/mnist/`, 1998.