Toward Certified Robustness Against Real-World Distribution Shifts

Haoze Wu1, Teruhiro Tagomori1,2, Alexander Robey3, Fengjun Yang3, Nikolai Matni3, George Pappas3, Hamed Hassani3, Corina Pasareanu4, Clark Barrett1
1 Stanford University, USA. 2 NRI Secure, Japan. 3 University of Pennsylvania, USA. 4 Carnegie Mellon University, USA.

Abstract—We consider the problem of certifying the robustness of deep neural networks against real-world distribution shifts. To do so, we bridge the gap between hand-crafted specifications and realistic deployment settings by considering a neural-symbolic verification framework in which generative models are trained to learn perturbations from data and specifications are defined with respect to the output of these learned models. A pervasive challenge arising from this setting is that although S-shaped activations (e.g., sigmoid, tanh) are common in the last layer of deep generative models, existing verifiers cannot tightly approximate S-shaped activations. To address this challenge, we propose a general meta-algorithm for handling S-shaped activations which leverages classical notions of counter-example-guided abstraction refinement. The key idea is to “lazily” refine the abstraction of S-shaped functions to exclude spurious counter-examples found in the previous abstraction, thus guaranteeing progress in the verification process while keeping the state-space small. For networks with sigmoid activations, we show that our technique outperforms state-of-the-art verifiers on certifying robustness against both canonical adversarial perturbations and numerous real-world distribution shifts. Furthermore, experiments on the MNIST and CIFAR-10 datasets show that distribution-shift-aware algorithms have significantly higher certified robustness against distribution shifts.

Index Terms—certified robustness, distribution shift, generative models, S-shaped activations, CEGAR

I. INTRODUCTION

Despite remarkable performance in various domains, it is well-known that deep neural networks (DNNs) are susceptible to seemingly innocuous variation in their input data. Indeed, recent studies have conclusively shown that DNNs are vulnerable to a diverse array of changes ranging from norm-bounded perturbations [1–7] to distribution shifts in weather conditions in perception tasks [8–12]. To address these concerns, there has been growing interest in using formal methods to obtain rigorous verification guarantees for neural networks with respect to particular specifications [13–51]. A key component of verification is devising specifications that accurately characterize the expected behavior of a DNN in realistic deployment settings. Designing such specifications is crucial for ensuring that the corresponding formal guarantees are meaningful and practically relevant.

By and large, the DNN verification community has focused on specifications described by simple analytical expressions. This line of work has resulted in a set of tools which cover specifications such as certifying the robustness of DNNs against norm-bounded perturbations [14, 31, 44, 52]. However, while such specifications are useful for certain applications, such as protecting against malicious security threats [53], there are many other applications where real-world distribution shifts, such as change in weather conditions, are more relevant, and these often cannot be described via a set of simple equations. While progress has been made toward broadening the range of specifications [54–57], it remains a crucial open challenge to narrow the gap between formal specifications and distribution shifts.

One promising approach for addressing this challenge involves incorporating neural network components in the specifications [57–59]. Such an approach has been used in the past to verify safety properties of neural network controllers [57] as well as robustness against continuous transformations between images [58]. More recently, Xie et al. [59] generalize this approach by proposing a specification language which can be used to specify complex specifications that are otherwise challenging to define. In this paper, we leverage and extend these insights to obtain a neural-symbolic (an integration of machine learning and formal reasoning) approach for verifying robustness against real-world distribution shifts. The key idea is to incorporate deep generative models that represent real-world distribution shifts [8, 9, 60, 61] in the formal specification.

To realize this idea, there remains one important technical challenge: all the previous work [57–59] assumes that both the neural networks being verified and the generative models are piecewise-linear. This assumption is made in order to leverage off-the-shelf neural network verifiers [57, 59], which focus on piecewise-linear activation functions such as ReLU. However, in practice the majority of (image) generative models [62–66] use S-shaped activation functions such as sigmoid and tanh in the output layer. While there are a few existing methods for verifying neural networks with sigmoidal activations [31, 34, 50, 52, 67], they all rely on a one-shot abstraction of the sigmoidal activations and suffer from lack of further progress if the verification fails on the abstraction. To bridge this gap and enable the use of a broad range of generative model architectures in the neural symbolic verification approach, we propose a novel abstraction-refinement algorithm.
for handling transcendental activation functions. We show that this innovation significantly boosts verification precision when compared to existing approaches.

Our neural-symbolic approach has the obvious limitation that the quality of the formal guarantee depends on the quality of the neural network used in the specification. However, the approach also possesses several pragmatically useful features that mitigate this limitation: if the verification fails, then a counter-example is produced; by examining the counter-example, we can determine whether it is a failure of the model being tested or a failure of the generative model. In either case, this gives us valuable information to improve the verification framework. For example, the failure of the generative model might point us to input regions where the generative model is under-trained and data augmentation is needed. On the other hand, if the verification succeeds, then the generative model represents a large class of inputs for which we know the model is robust.

Contributions. We summarize our contributions as follows:

- We describe a framework for verifying DNNs against real-world distribution shifts by incorporating deep generative models that capture distribution shifts---e.g., changes in weather conditions or lighting in perception tasks---as verification specifications.
- We propose a novel counter-example-guided abstraction refinement strategy for verifying networks with transcendental activation functions.
- We show that our verification techniques are significantly more precise than existing techniques on a range of challenging real-world distribution shifts on MNIST and CIFAR-10, as well as on canonical adversarial robustness benchmarks.

II. PROBLEM FORMULATION

In this section, we formally define the problem of verifying the robustness of DNN-based classifiers against real-world distribution shifts. The key step in our problem formulation is to propose a unification of logical specifications with deep generative models which capture distribution shifts.

Neural network classification. We consider classification tasks where the data consists of instances \( x \in \mathbb{X} \subseteq \mathbb{R}^d \) and corresponding labels \( y \in [k] := \{1, \ldots, k\} \). The goal of this task is to obtain a classifier \( C_f : \mathbb{R}^d \rightarrow [k] \) such that \( C_f \) can correctly predict the label \( y \) of each instance \( x \) for each \( (x, y) \) pair. In this work, we consider classifiers \( C_f \) defined by

\[
C_f(x) = \arg \max_{y \in [k]} f_j(x),
\]

where we take \( f : \mathbb{R}^{d_0} \rightarrow \mathbb{Y} \subseteq \mathbb{R}^{d_L} \) (with \( d_L = k \)) to be an \( L \)-layer feed-forward neural network with weights and biases \( W^{(i)} \in \mathbb{R}^{d_i \times d_{i-1}} \) and \( b^{(i)} \in \mathbb{R}^{d_i} \) for each \( i \in [L] \) respectively. More specifically, we let \( f(x) = n^{(L)}(x) \) and recursively define

\[
\begin{align*}
n^{(i)}(x) &= W^{(i)} \left( \hat{n}^{(i-1)}(x) + b^{(i)} \right), \\
\hat{n}^{(i)}(x) &= \rho \left( n^{(i)}(x) \right), \quad \text{and} \\
\hat{n}^{(0)}(x) &= x.
\end{align*}
\]

Here, \( \rho \) is a given activation function (e.g., ReLU, sigmoid, etc.) and \( n^{(i)} \) and \( \hat{n}^{(i)} \) represent the pre- and post-activation values of the \( i \)th layer of \( f \) respectively.

Perturbation sets and logical specifications. The goal of DNN verification is to determine whether or not a given logical specification regarding the behavior of a DNN holds in the classification setting described above. Throughout this work, we use the symbol \( \Phi \) to denote such logical specifications, which define relations between the input and output of a DNN. That is, given input and output properties \( \Phi_{\text{in}} \) and \( \Phi_{\text{out}} \) respectively, we express logical specifications \( \Phi \) in the following way:

\[
\Phi := \left( \Phi_{\text{in}}(x) \Rightarrow \Phi_{\text{out}}(y) \right).
\]

For example, given a fixed instance-label pair \( (x, y) \), the specification

\[
\Phi := \left( ||x - \bar{x}||_p \leq \epsilon \Rightarrow C_f(x) = y \right)
\]

captures the property of robustness against norm-bounded perturbations by checking whether all points in an \( \ell_p \)-norm ball centered at \( \bar{x} \) are classified by \( C_f \) as having the label \( y \).

Although the study of specifications such as (4) has resulted in numerous verification tools, there are many problems which cannot be described by this simple analytical model, including settings where data varies due to distribution shifts. For this reason, it is of fundamental interest to generalize such specifications to capture more general forms of variation in data. To do so, we consider abstract perturbation sets \( S(x) \), which following [9] are defined as “a set of instances that are considered to be equivalent to a fixed instance \( x \).” An example of an abstract perturbation set is illustrated in Figure 1b, wherein each instance in \( S(x) \) shows the same street sign with varying levels of snow. Ultimately, as in the case of norm-bounded robustness, the literature surrounding abstract perturbation sets has sought to train classifiers to predict the same output for each instance in \( S(x) \) [8, 9, 60].

Learning perturbation sets from data. Designing abstract perturbation sets \( S(x) \) which accurately capture realistic deployment settings is critical for providing meaningful guarantees. Recent advances in the generative modeling community have parameterized perturbation sets \( S(x) \) in the latent space \( \mathcal{Z} \) of a generative model \( G(x, z) \), where \( G \) takes as input an instance \( x \) and a latent variable \( z \in \mathcal{Z} \). Prominent among such works is [9], wherein the authors study the ability of conditional variational autoencoders (CVAEs) to capture shifts such as variation in lighting and weather conditions in images. In this
The majority of the verification literature has focused on a limited set of specifications, such as \( \ell_p \)-norm bounded perturbations, wherein perturbations can be defined by simple analytical expressions.

Most real-world perturbations cannot be described by simple analytical expressions. For example, obtaining a simple expression for a perturbation set \( S(x) \) that describes variation in snow would be challenging.

In particular, \( \mu(x) \) denotes a neural network that maps each instance \( x \) to the mean of a normal distribution over the latent space \( Z \) conditioned on \( x \). Similarly, \( \sigma(x) \) maps \( x \) to the standard deviation of this distribution in the latent space.

\(^1\)This notation is consistent with [9], wherein the authors use the same parameterization for conditional VAEs.

Under favorable optimization conditions, the authors of [9] prove that CVAEs satisfy two statistical properties which guarantee that the data belonging to learned perturbation sets in the form of (5) produce realistic approximations of the true distribution shift (c.f. Assumption 1 and Thms. 1 and 2 in [9]). In particular, the authors of [9] argue that if a learned perturbation set \( S(x) \) in the form of (5) has been trained such that the population-level loss is bounded by two absolute constants, then \( S(x) \) well-approximates the true distribution shift in the following sense: with high probability over the latent space, for any clean and perturbed pair \( (x, \tilde{x}) \) corresponding to a real distribution shift, there exists a latent code \( z \) such that \( ||z||_2 \leq \alpha \) and \( ||G(x, \mu(x)z - \sigma(x)) - \tilde{x}|| \leq \beta \) for two small constants \( \alpha \) and \( \beta \) depending on the CVAE loss. To further verify this theoretical evidence, we show that this framework successfully captures real-world shifts on MNIST and CIFAR-10 in Figure 2.

**Verifying robustness against learned distribution shifts.** To bridge the gap between formal verification methods and perturbation sets which accurately capture real-world distribution shifts, our approach in this paper is to incorporate perturbation sets parameterized by deep generative models into verification routines. We summarize this setting in the

---

\( S(x) = \{x + z : ||z|| \leq \delta\} \)

\( x_3 = G(x, z_3) \)

\( x_1 = G(x, z_1) \)

\( x_2 = G(x, z_2) \)
An abstraction of the sigmoid activation function.

In other words, our goal is to devise methods which verify whether a given classifier $C_f$ outputs the correct label $y$ for each instance in a perturbation set $S(x)$ parameterized by a generative model $G$.

III. TECHNICAL APPROACH AND CHALLENGES

The high-level idea of our approach is to consider the following equivalent specification to (6), wherein we absorb the generative model $G$ into the classifier $C$:

$$\Phi := (||z|| \leq \delta \implies C_{\phi}(z) = \overline{y})$$

In this expression, we define

$$Q_2(x) = (f \circ G)(x, \mu(x) + z\sigma(x))$$

to be the concatenation of the deep generative model $G$ with the DNN $f$. While this approach has clear parallels with verification schemes within the norm-bounded robustness literature, there is a fundamental technical challenge: state-of-the-art generative models typically use $S$-shaped activation functions (e.g., sigmoid, tanh) in the last layer to produce realistic data; however, the vast majority of the literature concerning DNN verification considers DNNs that are piece-wise linear functions. Therefore, existing methods for verification of generative models largely do not apply in this setting [57, 58].

Verification with S-shaped activations. In what follows, we describe the challenges inherent to verifying neural networks with S-shaped activations. For completeness, in the following definitions we provide a formal and general description of S-shaped activations, which will be crucial to our technical approach in this paper.

**Definition III.1** (Inflection point). A function $f : \mathbb{R} \to \mathbb{R}$ has an inflection point at $\eta$ iff it is twice differentiable at $\eta$, $f''(\eta) = 0$, and $f'$ changes sign as its argument increases through $\eta$.

**Definition III.2** (S-shaped function). An $S$-shaped function $\rho : \mathbb{R} \to \mathbb{R}$ is a bounded, twice differentiable function which has a non-negative derivative at each point and has exactly one inflection point.

In the wider verification literature, there are a handful of verification techniques that can handle $S$-shaped functions [18, 31, 34, 52, 67, 68]. Each of these methods is based on abstraction, which conservatively approximates the behavior of the neural network. We define abstraction formally in Sec. IV, but illustrate the key ideas here in Fig. 3, which shows the popular sigmoid activation function

$$\rho(x) = \frac{1}{1 + e^{-x}}. \quad (9)$$

Suppose the input $x$ is bounded between $l$ and $u$, the pre- and post-sigmoid values can be precisely described as $\mathcal{D} = \{(x, y) | y = \rho(x) \land l \leq x \leq u\}$. However, instead of using this precise representation, we could over-approximate $\mathcal{D}$ as $\mathcal{D'} = \{(x, y) | y \geq ax + b \land y \leq cx + d \land l \leq x \leq u\}$ (the yellow convex region in Fig. 3), where $ax + b$ and $cx + d$ are the two lines crossing the sigmoid function at $\{l, \rho(l)\}$ and $\{u, \rho(u)\}$. $\mathcal{D'}$ over-approximates $\mathcal{D}$ because the latter is a subset of the former. Abstraction-based methods typically over-approximate all non-linear connections in the neural network and check whether the specification holds on the over-approximation. The benefit is that reasoning about the (in this case linear) over-approximation is typically computationally easier than reasoning about the concrete representation, and moreover, if the specification holds on the abstraction, then it actually holds on the original network.

Previous work has studied different ways to over-approximate $S$-shaped activations, and there are trade-offs between how precise the over-approximation is and how efficient it is to reason about the over-approximation. For example, a piecewise-linear over-approximation can be more precise than the linear over-approximation (in Fig. 3), but reasoning about the former is more computationally challenging. However, whichever over-approximation one uses, all abstraction-based methods suffer from imprecision: if a counter-example to the specification is found on the over-approximation (which means the over-approximation violates the specification), we cannot conclude that the original network also violates the specification. This is because the counter-example may be spurious—inconsistent with the constraints imposed by the precise, unabtracted neural network (as shown in red in Fig. 3).

This spurious behavior demonstrates that there is a need for refinement of abstraction-based methods to improve the precision. For piecewise-linear activations, there is a natural...
refinement step: performing case analysis on the activation phases. However, when dealing with S-shaped activations, it is less clear how to perform this refinement, because if the refinement is performed too aggressively, the state space may explode and exceed the capacity of current verifiers. To address this technical challenge, we propose a counter-example guided refinement strategy for S-shaped activation functions which is based on the CEGAR approach [69]. Our main idea is to limit the scope of the refinement to the region around a specific counter-example. In the next section, we formally describe our proposed framework and we show that it can be extended to other transcendental activation functions (e.g., softmax).

IV. A CEGAR FRAMEWORK FOR S-SHAPED ACTIVATIONS

In this section, we formalize our meta-algorithm for precisely reasoning about DNNs with S-shaped activations, which is based on the CEGAR framework [70]. We first present the general framework and then discuss concrete instantiations of the sub-procedures.

Verification preliminaries. Our procedure operates on tuples of the form $M := (V, X, Y, \phi_{\text{aff}}, \phi_p)$. Here, $V$ is a set of real variables with $X \subseteq V$ and $\phi_{\text{aff}}$ and $\phi_p$ are sets of formulas over $V$ (when the context is clear, we also use $\phi_{\text{aff}}$ and $\phi_p$ to mean the conjunctions of the formulas in those sets). A variable assignment $\alpha : V \mapsto \mathbb{R}$ maps variables in $V$ to real values. We consider properties of the form $\Phi := (\Phi_{\text{in}}(X) \Rightarrow \Phi_{\text{out}}(Y))$, where $\Phi_{\text{in}}(X)$ and $\Phi_{\text{out}}(Y)$ are linear arithmetic formulas over $X$ and $Y$, and we say that $\Phi$ holds on $M$ if and only if the formula

$$\psi := \phi_{\text{aff}} \land \phi_p \land \Phi_{\text{in}}(X) \land \neg \Phi_{\text{out}}(Y)$$

is unsatisfiable. We use $M \models \Phi$ to denote that $\Phi$ holds (if $\psi$ is unsatisfiable), $M[\alpha] \models \neg \Phi$ to denote that $\Phi$ does not hold and is falsified if $\psi$ can be satisfied with assignment $\alpha$, and $M[\alpha] \models \Phi$ to denote that $\Phi$ is not falsified by $\alpha$ ($\alpha$ does not satisfy $\psi$). Given this notation, we define a sound abstraction as follows:

Definition IV.1 (Sound abstraction). Given a tuple $M := (V, X, Y, \phi_{\text{aff}}, \phi_p)$ and a property $\Phi := (\Phi_{\text{in}}(X) \Rightarrow \Phi_{\text{out}}(Y))$, we say that the tuple $M' := (V' \supseteq V, X, Y, \phi_{\text{aff}}', \phi_p')$ is a sound abstraction of $M$ if $M' \models \Phi$ implies that $M \models \Phi$.

Verifying DNNs. Given a DNN $f$, we construct a tuple $M_f$ as follows: for each layer $i$ in $f$, we let $v^{(i)}$ be a vector of $d_i$ variables representing the pre-activation values in layer $i$, and let $v^{(i)}$ be a similar vector representing the post-activation values in layer $i$. Let $v^{(0)}$ be a vector of $n_d$ variables from $V$ representing the inputs. Then, let $V$ be the union of all these variables, and let $X$ and $Y$ be the input and output variables, respectively; that is, $X$ consists of the variables in $v^{(0)}$, and $Y$ contains the variables in $v^{(L)}$. $\phi_{\text{aff}}$ and $\phi_p$ capture the affine and non-linear (i.e., activation) transformations in the neural network, respectively. In particular, for each layer $i$, $\phi_{\text{aff}}$ contains the formulas $v^{(i)} = W^{(i)}v^{(i-1)} + b^{(i)}$, and $\phi_p$ contains the formulas $v^{(i)} = \rho(v^{(i)})$.

Algorithm 1 presents a high-level CEGAR loop for checking whether $M \models \Phi$. It is parameterized by three functions. The $\text{Abstract}$ function produces an initial sound abstraction of $M$. The $\text{Prove}$ function checks whether $M' \models \Phi$. If so (i.e., the property $\Phi$ holds for $M'$), it returns with $\text{proven}$ set to true. Otherwise, it returns an assignment $\alpha$ which constitutes a counter-example. The final function is $\text{Refine}$, which takes $M$ and $M'$, the property $P$, and the counterexample $\alpha$ for $M'$ as inputs. Its job is to refine the abstraction until $\alpha$ is no longer a counter-example. If it succeeds, it returns a new sound abstraction $M''$. It fails if $\alpha$ is an actual counter-example for the original $M$. In this case, it sets the return value $\text{refined}$ to false. Throughout its execution, the algorithm maintains a sound abstraction of $M$ and checks whether the property $\Phi$ holds on the abstraction. If a counter-example $\alpha$ is found such that $M'[\alpha] \models \neg \Phi$, the algorithm uses it to refine the abstraction so that $\alpha$ is no longer a counter-example. The following theorem follows directly from Def. IV.1:

Theorem IV.2 (CEGAR is sound). Algorithm 1 returns true only if $M \models \Phi$.

A. Choice of the underlying verifier and initial abstraction

The $\text{Prove}$ function can be instantiated with an existing DNN verifier. The verifier is required to (1) handle piecewise-linear constraints; and (2) produce counter-examples. There are many existing verifiers that meet these requirements [14, 31, 44, 52]. To ensure that these two requirements are sufficient, we also require that $\phi_{\text{aff}}'$ and $\phi_p'$ only contain linear and piecewise-linear formulas.

The $\text{Abstract}$ function creates an initial abstraction. For simplicity, we assume that all piecewise-linear formulas are unchanged by the abstraction function. For S-shaped activations, we use piecewise-linear over-approximations. In principle, any sound piecewise-linear over-approximation of the S-shaped function could be used. One approach is to use a fine-grained over-approximation with piecewise-linear bounds [71]. While this approach can arbitrarily reduce over-approximation error, it might easily lead to an explosion of the state space when reasoning about generative models due to the large number of transcendental activations (equa...
is found by \( \in \mathbb{R} \) such that \( p \mid \mathcal{R} \) at \( \leq \) when the counter-example is \( \not\in \mathcal{R} \) and \( \not\in \mathcal{R} \). We assume that \( \phi'_{\text{aff}} \) includes upper and lower bounds for each variable \( v \) that is an input to an S-shaped function. In practice, bounds can be computed with bound-propagation techniques [31, 34, 35].

**Lemma IV.3.** Given an interval \( (l, u) \), an S-shaped function \( \rho \), and a point \( \{p, q\} \in \mathbb{R}^2 \), where \( p \in (l, u) \) and \( q \neq \rho(p) \), there exists a piecewise-linear function \( h : \mathbb{R} \rightarrow \mathbb{R} \) that 1) has two linear segments; 2) evaluates to \( \rho(p) \) at \( p \); and 3) separates \( \{p, q\} \) and \( \{(x, y) | x \in (l, u) \land y = \rho(x)\} \).

Leveraging this observation, given a point \( \{p, q\} = (\alpha(v), \alpha(\hat{v})) \), we can construct a piecewise-linear function \( h \) of the following form:

\[
h(x) = \rho(p) + \begin{cases} 
\beta(x - p) & \text{if } x \leq p \\
\gamma(x - p) & \text{if } x > p 
\end{cases}
\]

that separates the counter-example and the S-shaped function. If \( q > \rho(p) \), we add the formula \( \hat{v} \leq h(v) \) to the abstraction. And if \( q < \rho(p) \), we add \( \hat{v} \geq h(v) \).

The values for the slopes \( \beta \) and \( \gamma \) should ideally be chosen to minimize the over-approximation error while maintaining soundness. Additionally, they should be easily computable. Table I presents a general recipe for choosing \( \beta \) and \( \gamma \) when the spurious counter-example point is below the S-shaped function. Choosing \( \beta \) and \( \gamma \) when the counter-example is above the S-shaped function is symmetric (details are shown in App. A). \( \eta \) denotes the inflection point of the S-shaped function.

Note that in case 5, \( \beta \) is the same as \( \gamma \), meaning that a linear bound (the tangent line to \( \rho \) at \( p \)) suffices to exclude the counter-example. In terms of optimality, all but the \( \gamma \) value in case 1 and the \( \beta \) value in case 3 maximally reduce the over-approximation error among all valid slopes at \( (p, \rho(p)) \). In those two cases, linear or binary search techniques [34, 52] could be applied to compute better slopes, but the formulas shown give the best approximations we could find without using search.

**Lemma IV.4** (Soundness of slopes). Choosing \( \beta \) and \( \gamma \) using the recipe shown in Table I results in a piecewise-linear function \( h \) that satisfies the conditions of Lemma IV.3.

An instantiation of the \texttt{Refine} function for neural networks with S-shaped activation is shown in Alg. 2. It iterates through each S-shaped activation function. For the ones that are violated by the current assignment, the algorithm computes the slopes following the strategy outlined above with the \texttt{getSlopes} function and adds the corresponding piecewise-linear bounds (e.g., \( \hat{v} \leq h(v) \)) with the \texttt{addPLBound} function. Finally, we also allow the flexibility to terminate the refinement early with a customized \texttt{stopCondition} function. This is likely desirable in practice, as introducing a piecewise-linear bound for each violated activation might be too aggressive. Furthermore, adding a single piecewise-linear bound already suffices to exclude \( \alpha \). We use an adaptive stopping strategy where we allow at most \( m \) piecewise-linear bounds to be added in the first invocation of Alg. 2. And then, in each subsequent round, this number is increased by a factor \( k \). For our evaluation, below, we used \( m = 30 \) and \( k = 2 \), which were the values that performed best in an empirical analysis.

**Theorem IV.5** (Soundness of refinement). Given a sound abstraction \( M' \) of tuple \( M \), a property \( \Phi \), and a spurious counter-example \( \alpha \) s.t. \( M'_{\alpha} \models \neg \Phi \) and \( M_{\alpha} \models \Phi \). Alg. 2 produces a sound abstraction of \( M, M'' \), s.t. \( M''_{\alpha} \models \Phi \).

V. EXPERIMENTAL EVALUATION

In this section, we evaluate the performance of our proposed verification framework. We begin by investigating the effectiveness of our CEGAR-based approach on boosting the verification accuracy of existing approaches based on one-shot overapproximation. We evaluate on both robustness queries on real-world distribution shifts (§V-A) and existing benchmarks on robustness against norm-bounded perturbations [67, 72] (§V-C). Next, we benchmark the performance of our verifier on a range of challenging distribution shifts (§V-D). Finally, we use our method to show that robust training tends to result in higher levels of certified robustness against distribution shifts (§V-E).
We consider a diverse function with the Marabou neural network verification procedure (Marabou) on an encoding where each sigmoid is abstracted with the DeepPoly/CROWN linear bounds and the other parts are precisely encoded; and (3) DP+CEGAR, which is the CEGAR method proposed in this work. For each verifier, we perform a linear search for the largest perturbation bound each configuration can certify. Specifically, starting from $\delta = 0$, we repeatedly increase $\delta$ by 0.02 and check whether the configuration can prove robustness with respect to $S(x)$ within a given time budget (20 minutes). The process terminates when a verifier fails to prove or disprove a given specification.

For this experiment, we consider the shear distribution shift on MNIST and the fog distribution shift on the CIFAR-10 dataset (see Figure 2). All classifiers are trained using ERM. To provide a thorough evaluation, we consider several generator and classifier architectures; details can be found in App. D. Our results are enumerated in Table II, which shows the mean and standard deviation of the largest $\delta$ each configuration is able to prove for the first 100 correctly classified test images. We also report the average runtime on the largest $\delta$ proven robust by DP+BaB and DP+CEGAR, as well as the average number of abstraction refinement rounds by DP+CEGAR on those $\delta$ values. Across all configurations, our proposed technique effectively improves the verifiable perturbation bound with moderate runtime overhead. This suggests that the counter-example guided abstraction refinement scheme can successfully boost the precision when reasoning about sigmoid activations by leveraging existing verifiers.

### B. Effect of hyper-parameters $m$ and $k$

The refinement procedure (Alg. 2) has two numerical hyper-parameters, $m$ and $k$, which together control the number of new bounds introduced in a given refinement round (at round $i$, at most $m \times k^i$ new bounds are introduced). If too few new bounds are introduced, then there is not enough refinement, and the number of refinement rounds needed to prove the property increases. On the other hand, if too many bounds are introduced, we repeatedly increase $\delta$ and check that each verifier can certify. In our comparison, we consider three distinct configurations: (1) DP, which runs the DeepPoly/CROWN abstract interpretation procedure without any search; (2) DP+BaB, which runs a branch-and-bound

<table>
<thead>
<tr>
<th>Case</th>
<th>$l &lt; \eta, u &gt; \eta$</th>
<th>$l &lt; \eta, u &gt; \eta$</th>
<th>$l &lt; \eta, u &gt; \eta$</th>
<th>$l &gt; \eta \land u &lt; \eta$</th>
<th>$l &gt; \eta \land u &lt; \eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>$\rho'(p)$</td>
<td>$\rho'(p)$</td>
<td>$\rho'(p)$</td>
<td>$\rho'(p)$</td>
<td>$\rho'(p)$</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$\min(\rho'(l), \rho'(u))$</td>
<td>$\frac{\rho(p)-\rho(u)}{p-u}$</td>
<td>$\frac{\rho(p)-\rho(u)}{p-u}$</td>
<td>$\frac{\rho(p)-\rho(u)}{p-u}$</td>
<td>$\rho'(p)$</td>
</tr>
</tbody>
</table>

**TABLE I:** Slopes for the piece-wise linear abstraction refinement. The figures illustrate the refinement on the sigmoid function.
introduced, then the time required by the solver to check the abstraction might unnecessarily increase, resulting in timeouts.

To study the effect of \( m \) and \( k \) more closely, we evaluate the runtime performance of \( \text{DP+CEGAR} \) instantiated with different combinations of \( m \) and \( k \). In particular, we run all combinations of \( m \in \{10, 30, 50\} \) and \( k \in \{1.5, 2\} \) on the first 10 verification instances for the first two generator-classifier pairs in Table II—(\( \text{MLP}\_\text{GEN1}, \text{MLP}\_\text{CLS1} \)) and (\( \text{CONV}\_\text{GEN1}, \text{CONV}\_\text{CLS1} \)). The perturbation bound is the largest \( \delta \) proven robust for each instance.

Table III shows the number of solved instances, the average runtime on solved instances, and the average number of refinement rounds on solved instances for the same value of \( k \), the number of refinement rounds on solved instances consistently decreases as \( m \) increases. This is expected, because refinements are performed more eagerly as \( m \) increases. However, the decrease in the number of refinement rounds does not necessarily imply improvements in performance. For example, \((50, 2)\) solves one fewer MNIST benchmark than \((30, 2)\). On the other hand, if the strategy is too “lazy” (e.g., \( m \) is too small), the increased number of refinement rounds can also result in runtime overhead. For example, on the CIFAR10 benchmarks, when \( k = 1.5 \), the average runtime decreases as \( m \) increases, even though all three configurations solve the same number of instances.

Overall, this study suggests that the optimal values of \( m \) and \( k \) vary across different benchmarks, and exploring adaptive heuristics for choosing these hyper-parameters is a promising direction for boosting the runtime performance of the proposed algorithm.

C. Further evaluation of CEGAR on adversarial robustness benchmarks.

To better understand the effectiveness of our abstraction-refinement technique on boosting the verification accuracy over one-shot approximation, we consider a different initial abstraction, PRIMA [67], a more recently proposed abstraction that considers convex relaxations over groups of activation functions and is empirically more precise than DeepPoly/CROWN. In particular, we focus on the same sigmoid benchmarks used in [67]. We use the PRIMA implementation in the artifact associated with the paper4 and run a configuration PRIMA+CEGAR which is the same as \( \text{DP+CEGAR} \), except that we run PRIMA instead of DeepPoly/CROWN for the initial abstraction. Each job is given 16 threads and a 30 minute wall-clock time limit. Table IV shows the number of verified instances and the average runtime on verified instances for the two configurations. Our configuration is able to consistently solve more instances with only a moderate increase in average solving time. This suggests that the meta-algorithm that we propose can leverage the tighter bounds derived by existing abstraction-based methods and boost the state of the art in verification accuracy on sigmoid-based networks.

### Table II: Evaluation results of three solver configurations.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( m )</th>
<th>( k )</th>
<th># solved</th>
<th>time(s)</th>
<th># ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>10</td>
<td>1.5</td>
<td>10</td>
<td>139.5</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>1.5</td>
<td>8</td>
<td>52.4</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1.5</td>
<td>7</td>
<td>39.8</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2</td>
<td>10</td>
<td>113.6</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>2</td>
<td>10</td>
<td>121.0</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>2</td>
<td>9</td>
<td>64.6</td>
<td>0.9</td>
</tr>
<tr>
<td>CIFAR</td>
<td>10</td>
<td>1.5</td>
<td>9</td>
<td>77.0</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>1.5</td>
<td>9</td>
<td>60.4</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1.5</td>
<td>9</td>
<td>38.2</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2</td>
<td>9</td>
<td>55.3</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>2</td>
<td>10</td>
<td>100.7</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>2</td>
<td>10</td>
<td>82.2</td>
<td>1.6</td>
</tr>
</tbody>
</table>

### Table III: Effect of \( m \) and \( k \) on the runtime performance of CEGAR.

### Table IV: Comparison with PRIMA [67] on the same benchmarks used in [67].

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc.</th>
<th>( \epsilon )</th>
<th>PRIMA time(s)</th>
<th>PRIMA+CEGAR time(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6x100</td>
<td>99</td>
<td>0.015</td>
<td>52</td>
<td>65</td>
</tr>
<tr>
<td>9x100</td>
<td>99</td>
<td>0.015</td>
<td>57</td>
<td>66</td>
</tr>
<tr>
<td>6x200</td>
<td>99</td>
<td>0.012</td>
<td>65</td>
<td>75</td>
</tr>
<tr>
<td>ConvSmall</td>
<td>99</td>
<td>0.014</td>
<td>56</td>
<td>63</td>
</tr>
</tbody>
</table>

We have also evaluated our techniques on the sigmoid benchmarks used in VNN-COMP 2021 (there are no sigmoid benchmarks in VNN-COMP 2022), and we are also able to boost the verification precision over other SoTA solvers such as \( \alpha = \beta \)-CROWN and VeriNet [52] with our approach. Details can be found in App. E.
Robustness of ERM against different perturbations.

For instance, for the shear increases to 0.2. Information like this could be considered: 0.1, 0.2, and 0.5. The architectures we consider are: MLP\_GEN2 and MLP\_CLS3. For CIFAR-10 we use CONV\_GEN2 and CONV\_CLS2. The verification results are shown in Figure 4. The “robust” columns show the number of instances that our verification procedure is able to certify within a 20 minute timeout. As one would expect, the robustness of each classifier deteriorates as the perturbation budget \( \delta \) increases. For instance, for the shear transformation, the classifier is robust on 76 out of the 100 instances when \( \delta = 0.1 \), but is only certified robust on 4 instances when \( \delta \) increases to 0.2. Information like this could help system developers to identify perturbation classes for which the network is especially vulnerable and potentially retrain the network accordingly.

### D. Benchmarking our approach on an array of real-world distribution shifts

We next use our proposed verification procedure to evaluate the robustness of classifiers trained using ERM against a wide range of distribution shifts. We select the first 100 correctly classified test images from the respective dataset for each perturbation and verify the robustness of the classifier against the perturbation set. Three values of the perturbation variable \( \delta \) are considered: 0.1, 0.2, and 0.5. The architectures we consider for MNIST are MLP\_GEN2 and MLP\_CLS3. For CIFAR-10 we use CONV\_GEN2 and CONV\_CLS2. The verification results are shown in Figure 4. The “robust” columns show the number of instances that our verification procedure is able to certify within a 20 minute timeout. As one would expect, the robustness of each classifier deteriorates as the perturbation budget \( \delta \) increases. For instance, for the shear transformation, the classifier is robust on 76 out of the 100 instances when \( \delta = 0.1 \), but is only certified robust on 4 instances when \( \delta \) increases to 0.2. Information like this could help system developers to identify perturbation classes for which the network is especially vulnerable and potentially retrain the network accordingly.

### E. Verification for various robust training algorithms

Finally, we compare the robustness and accuracy of classifiers trained using ERM, IRM, PGD, and MDA against the shear distribution shifts on the MNIST dataset. To this end, we measure the accuracy on the entire test set under the learned perturbation generative models. For each classifier, we then select the first 500 correctly classified images in its dataset and verify the targeted robustness of the classifier against the perturbation. The architectures we use are MLP\_GEN2 and MLP\_CLS3.

Accuracy and robustness results are presented in Figure 5. Interestingly, MDA, which is perturbation-aware, outperforms the other three perturbation-agnostic training methods, on both test accuracy and robustness, suggesting that knowing what type of perturbation to anticipate is highly useful. Notice that accuracy on the perturbation set is not necessarily a good proxy for robustness: while the IRM-trained classifier has similar accuracy as the PGD-trained classifier, the former is significantly less robust on the perturbation set with \( \delta = 0.1 \). This further supports the need for including formal verification in the systematic evaluation of neural networks and training algorithms.

### VI. Related Work

#### Beyond norm-bounded perturbations

While the literature concerning DNN verification has predominantly focused on robustness against norm-bounded perturbations, some work has considered other forms of robustness, e.g., against geometric transformations of data [54–56]. However, the perturbations considered are hand-crafted and can be analytically defined by simple models. In contrast, our goal is to verify against real-world distribution shifts that are defined via the output set of a generative model. Our approach also complements recent work which has sought to incorporate neural symbolic components into formal specifications [59]. Our work differs from [59] in two ways. Firstly, Xie et al. use classifiers and regressors in their neural specifications, while we use generative models to express perturbation sets in our specifications. Secondly, while Xie et al. make black-box use of existing verifiers, we propose a new verification scheme for analyzing networks with sigmoid activations. In general, we believe the idea of leveraging neural components for specification is very promising and should be explored further.

#### Existing verification approaches

Existing DNN verification algorithms broadly fall into one of two categories: search-based methods [13–30] and abstraction-based methods [31–51]. While several existing solvers can handle sigmoid activation functions [18, 31, 34, 52, 67], they rely on one-shot abstraction and lack a refinement scheme for continuous progress. On the other hand, a separate line of work has shown that verifying DNNs containing a single layer of logistic activations is decidable [77], but the decision procedure proposed in this work is computationally prohibitive. To overcome these limitations, we propose a meta-algorithm inspired by counter-example-guided abstraction refinement [69] that leverages existing verifiers to solve increasingly refined abstractions. We notice a concurrent work [78] on formal reasoning of sigmoidal activations which is made available a week before the deadline. The technique is orthogonal to our approach as it again performs one-shot over-approximation of sigmoidal activations.

CEGAR [69, 79] is a well-known technique. Our contribution in this area lies in investigating the choices of its parameters (i.e., Abstract, Prove, and Refine) when it comes to verifying neural networks with S-shaped activations.

#### Verification against distribution shifts

The authors of [9] also considered the task of evaluating DNN robustness to real-world distribution shifts; in particular, the approach used

---

**Table 1:**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Perturbation</th>
<th>( \delta = 0.1 ) robust time(s)</th>
<th>( \delta = 0.2 ) robust time(s)</th>
<th>( \delta = 0.5 ) robust time(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>brightness</td>
<td>99.3</td>
<td>95.5</td>
<td>89.5</td>
</tr>
<tr>
<td></td>
<td>rotation</td>
<td>51.2</td>
<td>80.1</td>
<td>177.9</td>
</tr>
<tr>
<td></td>
<td>shear</td>
<td>76.2</td>
<td>102.6</td>
<td>135.6</td>
</tr>
<tr>
<td></td>
<td>contrast</td>
<td>90.2</td>
<td>11.1</td>
<td>182.0</td>
</tr>
<tr>
<td></td>
<td>scale</td>
<td>95.8</td>
<td>30.8</td>
<td>122.7</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>brightness</td>
<td>97.2</td>
<td>96.5</td>
<td>86.5</td>
</tr>
<tr>
<td></td>
<td>contrast</td>
<td>97.2</td>
<td>4.6</td>
<td>77.0</td>
</tr>
<tr>
<td></td>
<td>fog</td>
<td>84.3</td>
<td>69.1</td>
<td>1125.0</td>
</tr>
<tr>
<td></td>
<td>gaussian-blur</td>
<td>100.0</td>
<td>3.9</td>
<td>94.0</td>
</tr>
</tbody>
</table>

**Figure 4:** Robustness of ERM against different perturbations.

**Figure 5:** Test set accuracy and verification accuracy.

---

1https://dl.acm.org/doi/10.1145/3462308/full/
in [9] relies on randomized smoothing [80]. This scheme provides probabilistic guarantees on robustness, whereas our approach (as well as the aforementioned approaches) provides deterministic guarantees. In a separate line of work, several authors have sought to perform verification of deep generative models [57, 58]. However, each of these works assumes that generative models are piece-wise linear functions, which precludes the use of state-of-the-art models.

VII. CONCLUSION

In this paper, we presented a framework for certifying robustness against real-world distribution shifts. We proposed using provably trained deep generative models to define formal specifications and a new abstraction-refinement algorithm for verifying them. Experiments show that our method can certify against larger perturbation sets than previous techniques.

Limitations. We now discuss some limitations of our framework. First, like many verification tools, the classifier architectures that our approach can verify are smaller than popular architectures such as ResNet [81] and DenseNet [82]. This stems from the limited capacity of existing DNN verification tools for piece-wise-linear functions, which we invoke in the CEGAR loop. Given the rapid growth of the DNN verification community, we are optimistic that the scalability of verifiers will continue to grow rapidly, enabling their use on larger and larger networks. Another limitation is that the quality of our neural symbolic specification is determined by how well the generative model captures real-world distribution shifts. The mismatch between formal specification and reality is in fact common (and often unavoidable) in formal verification. And while [9] shows that under favorable conditions, CVAEs can capture distribution shifts, these assumptions may not hold in practice. For this reason, we envision that in addition to these theoretical results, a necessary future direction will be to involve humans in the verification loop to validate the shifts captured by generative models and the produced counterexamples. This resembles how verification teams work closely with product teams to continually re-evaluate and adjust the specifications in existing industrial settings. In general, we believe that closing this verification loop (verify, debug, verify again, etc.) is a very interesting future research direction. Finally, when applying our techniques in real industrial settings, another challenge is that collecting training data corresponding to new distribution shifts may be costly (e.g., collecting the same street view under different times of day). However, this cost may be justified in safety-critical domains.

REFERENCES


APPENDIX A
CHOICES OF SLOPES (CONT.)

We present in Table V the general recipe for choosing $\beta$ and $\gamma$ in the case when the violation point is above the S-shaped function.

APPENDIX B
PROOFS

Proof. Theorem IV.2. Alg. 1 returns true only if the property holds on a sound abstraction of $M$, which following Def. IV.1 means the property holds on $M$.

Proof. Lemma IV.3. This can be proved by construction using the $\beta$ and $\gamma$ values in Table I and Table V. We next prove that those choices are sound in Lemma IV.4.

Before proving Lemma IV.4, we first state the following definitions and facts. 5

Definition B.1 (Tagent line). The tangent line at $a$ to the function $f$, denoted by $\text{TanLine}_{f,a}(x)$, is defined as: $\text{TanLine}_{f,a}(x) = f(a) + f'(a)(x-a)$.

Definition B.2 (Secant line). Given $a, b \in \mathbb{R}$, the secant line at $[a, b]$ to a function $f$, denoted by $\text{SecLine}_{f,a,b}(x)$, is defined as: $\text{SecLine}_{f,a,b}(x) = \frac{f(b) - f(a)}{b-a} * (x-a) + f(a)$.

Proposition B.3. Let $f$ be a twice differentiable univariate function. If $f''(x) \geq 0$ for all $x \in [l, u]$, then for all $a, x \in [l, u]$, $\text{TanLine}_{f,a}(x) \leq f(x)$, and for all $a, b, x \in [l, u]$, where $a < b$ and $a \leq x \leq b$, $\text{SecLine}_{f,a,b}(x) \geq f(x)$.

Proposition B.4. Let $f$ be a twice differentiable univariate function. If $f''(x) \leq 0$ for all $x \in [l, u]$, then for all $a, x \in [l, u]$, $\text{TanLine}_{f,a}(x) \geq f(x)$, and for all $a, b, x \in [l, u]$, where $a < b$ and $a \leq x \leq b$, $\text{SecLine}_{f,a,b}(x) \leq f(x)$.

Proposition B.5. Let $f$ be a univariate function, differentiable with non-negative derivative on $[l, u]$. If $\gamma \leq f'(x)$ for all $x \in [l, u]$, then $f(l) + \gamma(x-l) \leq f(x)$ for all $x \in [l, u]$. We prove this is true for each case in Table. I.

Case 1: The segment corresponding to $\beta$ is $\text{TanLine}_{p,p}$, and Cond. (a) holds by Prop. B.3. On the other hand, the choice $\gamma$ is such that $\gamma \leq \rho'(x)$ for all $x \in [p, u]$. Thus, Cond. (b) holds by Prop. B.5.

Case 2: The segment corresponding to $\beta$ is $\text{TanLine}_{p,p}$, so Cond. (a) holds by Prop. B.3. The segment corresponding to $\gamma$ is $\text{SecLine}_{p,p,a}$, so Cond. (b) holds by Prop. B.4.

Case 3: For Cond. (a), we further break it into 2 cases: $x \leq \eta$ and $x > \eta$. In the former case, the line $\rho(p)+\beta(x-p)$ is below the line $\rho(l)+\min(\rho'(l), \rho'(u))(x-l)$, which

5These are partially adapted from [79].
Theorem IV.5. which is a counter-example with respect to $\rho$ is below $\rho$. When $x > \eta$, $\rho(p) + \beta(x - p)$ is below the secant line $\text{SecLine}_{\rho,p,p}$, which by Prop. B.4 is below $\rho$. On the other hand, the segment corresponding to $\gamma$ is $\text{SecLine}_{\rho,p,u}$, so Cond. (b) holds by Prop. B.4.

- **Case 4**: The segments are both secant lines, $\text{SecLine}_{\rho,t,p}$ and $\text{SecLine}_{\rho,p,u}$, and thus the conditions hold by Prop. B.4.

- **Case 5**: The segments are both tangent lines, $\text{TanLine}_{\rho,p}$, and thus the conditions hold by Prop. B.3.

The proof for the cases shown in Table. V is analogous. □

**Proof. Theorem IV.5.** We can prove the soundness of $M''$ by induction on the number of invocations of the ADDPLBOUND method. If it is never invoked, then $M'' = M'$ which is a sound abstraction. In the inductive case, it follows from Lemma IV.4 that adding an additional piecewise-linear bound does not exclude variable assignments that respect the precise sigmoid function. On the other hand, when $M[\alpha] = \Phi$, the ADDPLBOUND method will be invoked at least once, which precludes $\alpha$ as a counter-example with respect to $M''$. That is, $M''[\alpha] = \Phi$. □

**APPENDIX C**

**ENCODING PIECE-WISE LINEAR REFINEMENT USING LEAKYRELU**

We observe that it is possible to encode the piecewise-linear bounds that we add during abstraction refinement using LeakyReLU functions. While we do not leverage this fact in this paper, we lay out the reduction to LeakyReLU in this section to show how future work using verification tools supporting LeakyRelUs could benefit.

A LeakyReLU $r_\alpha$ is a piecewise linear function with two linear segments:

$$r_\alpha(x) = \begin{cases} \alpha \cdot x & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

where $\alpha \geq 0$ is a hyper-parameter.

Given a piecewise linear function with two linear segments:

$$h(x) - \rho(p) = \begin{cases} \beta(x - p) & \text{if } x \leq p \\ \gamma(x - p) & \text{if } x > p \end{cases}$$

We can rewrite $h$ as the following:

$$h(x) = \gamma \cdot r_\alpha(x - p) + \rho(p), \text{ where } \alpha := \frac{\beta}{\gamma}$$

Note that the $\alpha$ value is always valid (i.e., $\alpha \geq 0$) because we always choose both $\beta$ and $\gamma$ to be positive. This means that we can potentially encode the piecewise linear bounds as affine and leaky relu layers. For example, the piecewise linear upper bound $y \leq h(x)$ for a sigmoid $y = \rho(x)$ can be encoded as

$$\begin{align*}
    a_1 &= x - p \\
    a_2 &= r_\alpha(a_1) \\
    a_3 &= a_2 + \rho(p) \\
    y &= a_3 + a_4 \\
    a_4 &\leq 0,
\end{align*}$$

where $a_1, a_2, a_3, a_4$ are fresh auxiliary variables. Eqs. a) and c) can be modeled by feed-forward layers. Eq. b) can be modeled by a leaky relu layer. If we treat $a_3$ as an input variable to the neural network, Eq. d) can be modeled as a residual connection. This suggests that we could, in principle, express the abstraction as an actual piecewise-linear neural network (with bounds on the input variables (e.g., $a_1$), making it possible to leverage verifiers built on top of neural network software platforms such as Tensorflow or Pytorch.

**APPENDIX D**

**DETAILS ON TRAINING AND CVAES**

**A. Dataset**

We consider the well-known MNIST and CIFAR-10 datasets. The MNIST dataset contains 70,000 grayscale images of handwritten digits with dimensions $28 \times 28$, where we used 60,000 images for training and held 10,000 for testing. The CIFAR-10 dataset contains 60,000 colored images of 10 classes with dimensions $3 \times 32 \times 32$, where we used 50,000 images for training and held 10,000 for testing.

To perturb the images, we adapt the perturbations implemented in [10]. When training and testing the models, we

---

Table V: Slopes for the piece-wise linear abstraction refinement.

<table>
<thead>
<tr>
<th>Case</th>
<th>Condition</th>
<th>Slope $\beta$</th>
<th>Slope $\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>$l &lt; \eta, u &gt; \eta$</td>
<td>$\rho'(p) &gt; 0$</td>
<td>$\rho'(p)$</td>
</tr>
<tr>
<td>7</td>
<td>$l &lt; \eta, u &gt; \eta$</td>
<td>$\rho'(p) = 0$</td>
<td>$\rho'(p)$</td>
</tr>
<tr>
<td>8</td>
<td>$l &lt; \eta, u &gt; \eta$</td>
<td>$\rho'(p) &lt; 0$</td>
<td>$\rho'(p)$</td>
</tr>
<tr>
<td>9</td>
<td>$l &gt; \eta, u &lt; \eta$</td>
<td>$\rho'(p) \leq 0$</td>
<td>$\rho'(p)$</td>
</tr>
<tr>
<td>10</td>
<td>$l &gt; \eta, u &lt; \eta$</td>
<td>$\rho'(p) &gt; 0$</td>
<td>$\rho'(p)$</td>
</tr>
</tbody>
</table>

https://github.com/hendrycks/robustness/blob/master/ImageNet-C/create_c/Makefile_cifar_c.py

550

Authorized licensed use limited to: Stanford University. Downloaded on September 03,2023 at 00:00:20 UTC from IEEE Xplore. Restrictions apply.
sample images from the dataset and randomly perturb each image with a strength parameter \( c \) that is sampled uniformly from the ranges given in Table VI.

**TABLE VI:** Perturbation range in the training data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Perturbation</th>
<th>Range of ( c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>brightness</td>
<td>[0, 5]</td>
</tr>
<tr>
<td></td>
<td>rotation</td>
<td>[-60, 60]</td>
</tr>
<tr>
<td></td>
<td>gaussian blur</td>
<td>[0, 2, 1, 0]</td>
</tr>
<tr>
<td></td>
<td>shear</td>
<td>[0.0, 0.4]</td>
</tr>
<tr>
<td></td>
<td>contrast</td>
<td>[1.0, 5.0]</td>
</tr>
<tr>
<td></td>
<td>translate</td>
<td>[0.7, 0.9]</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>brightness</td>
<td>[0.05, 3]</td>
</tr>
<tr>
<td></td>
<td>contrast</td>
<td>[1.5, 75]</td>
</tr>
<tr>
<td></td>
<td>fog</td>
<td>[2, 1.5, 1.75, 3]</td>
</tr>
<tr>
<td></td>
<td>gaussian blur</td>
<td>[4, 1]</td>
</tr>
</tbody>
</table>

**B. Architecture**

On each dataset, we train a conditional variational encoder (CVAE) with three components: prior network, encoder network, and decoder network (generator). We also train a set of classifiers. In this section, we detail the architecture of these networks. The architectures for the MNIST networks are shown in Tables VII–XIII. Those of the CIFAR networks are shown in Tables XIV–XIX. The output layers of the generators use sigmoid activation functions. All hidden-layers use ReLU activation functions.

**TABLE VII:** Prior

<table>
<thead>
<tr>
<th>Type</th>
<th>Parameters/Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>28 x 28</td>
</tr>
<tr>
<td>Dense</td>
<td>784 x 1</td>
</tr>
<tr>
<td>Dense</td>
<td>300 x 1</td>
</tr>
<tr>
<td>Dense</td>
<td>8 x 2</td>
</tr>
</tbody>
</table>

**TABLE VIII:** Encoder

<table>
<thead>
<tr>
<th>Type</th>
<th>Parameters/Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>28 x 28 x 2</td>
</tr>
<tr>
<td>Dense</td>
<td>784 x 1</td>
</tr>
<tr>
<td>Dense</td>
<td>300 x 1</td>
</tr>
<tr>
<td>Dense</td>
<td>8 x 2</td>
</tr>
</tbody>
</table>

**TABLE IX:** MLP_GEN1

<table>
<thead>
<tr>
<th>Type</th>
<th>Parameters/Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>28 x 28 x 8</td>
</tr>
<tr>
<td>Dense</td>
<td>784 x 1</td>
</tr>
<tr>
<td>Dense</td>
<td>300 x 1</td>
</tr>
</tbody>
</table>

**TABLE X:** MLP_GEN2

<table>
<thead>
<tr>
<th>Type</th>
<th>Parameters/Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>28 x 28 x 8</td>
</tr>
<tr>
<td>Dense</td>
<td>784 x 1</td>
</tr>
<tr>
<td>Dense</td>
<td>300 x 1</td>
</tr>
</tbody>
</table>

**TABLE XI:** MLP_CLS1

<table>
<thead>
<tr>
<th>Type</th>
<th>Parameters/Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>28 x 28</td>
</tr>
<tr>
<td>Dense</td>
<td>32 x 1</td>
</tr>
<tr>
<td>Dense</td>
<td>32 x 1</td>
</tr>
<tr>
<td>Dense</td>
<td>10 x 1</td>
</tr>
</tbody>
</table>

**TABLE XII:** MLP_CLS2

<table>
<thead>
<tr>
<th>Type</th>
<th>Parameters/Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>28 x 28</td>
</tr>
<tr>
<td>Dense</td>
<td>64 x 1</td>
</tr>
<tr>
<td>Dense</td>
<td>32 x 1</td>
</tr>
<tr>
<td>Dense</td>
<td>10 x 1</td>
</tr>
</tbody>
</table>

**C. Optimization**

We implement our models and training in PyTorch. The CVAE implementation is adapted from that in [9]. On both datasets, we trained our CVAE networks for 150 epochs using the ADAM optimizer with a learning rate of \( 10^{-4} \) and forgetting factors of 0.9 and 0.999. In addition, we applied cosine annealing learning rate scheduling. Similar to [3], we increase \( \beta \) linearly from \( \beta = 0 \) at epoch 1 to \( \beta = 0.01 \) at epoch 40, before keeping \( \beta = 0.01 \) for the remaining epochs. We use a batch size of 256.

The ERM classifiers on the MNIST dataset are trained with the ADAM optimizer with a learning rate of \( 10^{-3} \) for 20 epochs. The classifiers for the CIFAR-10 dataset are trained with the ADAM optimizer with learning rate \( 10^{-3} \) for 200 epochs. The classifiers in Sec. V-E are also all trained with the ADAM optimizer with a learning rate of \( 10^{-3} \) for 20 epochs. For PGD, we use a step size of \( \alpha = 0.1 \), a perturbation budget of \( \epsilon = 0.3 \), and we use 7 steps of projected gradient ascent. For IRM, we use a small held-out validation set to select \( \lambda \in \{0.1, 1, 10, 100, 1000\} \). For MDA, we use a step size of \( \alpha = 0.1 \), a perturbation budget of \( \epsilon = 1.0 \), and we use 10 steps of projected gradient ascent.

**D. Computing resources**

The classifiers used in Section V-E were trained using a single NVIDIA RTX 5000 GPU. The other networks were trained using 8 AMD Ryzen 7 2700 Eight-Core Processors.

**APPENDIX E**

**EVALUATION ON VNN-COMP-21 BENCHMARKS**

We also evaluate our techniques on the 36 sigmoid benchmarks used in VNN-COMP-2021. We exclude the benchmark
where a counter-example can be found using the PGD attack and evaluate on the remaining 35 benchmarks. In particular, we run a sequential portfolio approach where we first attempt to solve the query with \( \alpha - \beta \)-CROWN [18, 34, 44] (competition version), and if the problem is not solved, we run PRIMA+CEGAR. Table IV shows the results. As a point of comparison, we also report the numbers of the top three performing tools [18, 31, 44, 52, 67] during VNN-COMP-21 on these benchmarks. \(^7\) While \( \alpha - \beta \)-CROWN is already able to solve 29 of the 35 benchmarks, with the abstraction refinement scheme, we are able to solve 1 additional benchmark. We note that during the competition, \( \alpha - \beta \)-crown did not exhaust the 5 minute per-instance timeout on any of these benchmarks.\(^8\) This suggests that the solver was not able to make further progress once the analysis is inconclusive on the one-shot abstraction of the sigmoid activations. On the other hand, our technique provides a viable way to make continuous progress if the one-shot verification attempt fails.

### APPENDIX F

**Evaluation of an Eager Refinement Strategy**

We also compare the lazy abstraction refinement strategy with an eager approach where piecewise-linear bounds are added for each sigmoid from the beginning instead of added lazily as guided by counter-examples. In particular, we attempt to add one piecewise-linear upper-bound and one piecewise-linear lower-bound, each with \( K \) linear segments, for each sigmoid activation function. The segment points are evenly distributed along the x-axis. We evaluate on the same MNIST benchmarks as in Table II, using \( K = 2 \) and \( K = 3 \). The results are shown in Table XXI. While the two strategies are still able to improve upon the perturbation bounds found by the pure abstract-interpretation-based approach \( DP \), the means of the largest certified \( \delta \) values for the eager approach are significantly smaller than those of the CEGAR-based configuration we propose. Interestingly, while \( K=3 \) uses a finer-grained over-approximation compared with \( K=2 \), the former only improves on one of the six benchmark sets. This suggests that the finer-grained abstraction increases the overhead to the solver and is not particularly effective at excluding spurious counter-examples on the set of benchmarks that we consider, which supports the need for a more informed abstraction refinement strategy such as the one we propose.

### APPENDIX G

**Licenses**

The MNIST and CIFAR-10 datasets are under the MIT License (MIT). The Marabou verification tool is under the terms of the modified BSD license (https://github.com/NeuralNetworkVerification/Marabou/blob/master/COPYING).

---

\(^7\)https://arxiv.org/abs/2109.00498

\(^8\)https://github.com/stanleybak/vnncomp2021_results/blob/main/results_csv/a-b-CROWN.csv
TABLE XX: Comparison on the VNN-COMP-21 benchmarks

<table>
<thead>
<tr>
<th>Model</th>
<th># Bench.</th>
<th>α</th>
<th>β-CROWN</th>
<th>VenNet</th>
<th>ERAN</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>robust</td>
<td>time(s)</td>
<td>robust</td>
<td>time(s)</td>
</tr>
<tr>
<td>6x200</td>
<td>35</td>
<td></td>
<td>29</td>
<td>12.9</td>
<td>20</td>
<td>2.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gen. Class.</th>
<th>K=2</th>
<th></th>
<th>K=3</th>
<th></th>
<th>DP+CEGAR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>δ</td>
<td>time(s)</td>
<td>δ</td>
<td>time(s)</td>
<td>δ</td>
</tr>
<tr>
<td>MNIST</td>
<td>MLP_GEN₁</td>
<td>0.137 ± 0.043</td>
<td>88.9</td>
<td>0.137 ± 0.042</td>
<td>109.8</td>
<td><strong>0.157</strong> ± 0.057</td>
</tr>
<tr>
<td>MLP_GEN₂</td>
<td>MLP_CLS₁</td>
<td>0.109 ± 0.031</td>
<td>114.5</td>
<td>0.109 ± 0.031</td>
<td>199.0</td>
<td><strong>0.118</strong> ± 0.049</td>
</tr>
<tr>
<td>MLP_GEN₃</td>
<td>MLP_CLS₂</td>
<td>0.126 ± 0.045</td>
<td>64.0</td>
<td>0.129 ± 0.044</td>
<td>95.9</td>
<td><strong>0.15</strong> ± 0.059</td>
</tr>
<tr>
<td>MLP_GEN₄</td>
<td>MLP_CLS₃</td>
<td>0.108 ± 0.038</td>
<td>150.0</td>
<td>0.106 ± 0.036</td>
<td>133.8</td>
<td><strong>0.121</strong> ± 0.049</td>
</tr>
</tbody>
</table>

**TABLE XXI**: Evaluation results of the eager approach. We also report again the results of DP+CEGAR, which is the same as Table II.