

---

# Provably Bounding Neural Network Preimages

---

**Suhas Kotha\***  
Carnegie Mellon  
suhas@andrew.cmu.edu

**Christopher Brix\***  
RWTH Aachen  
brix@cs.rwth-aachen.de

**Zico Kolter**  
Carnegie Mellon  
Bosch Center for AI  
zkolter@cs.cmu.edu

**Krishnamurthy (Dj) Dvijotham†**  
Google DeepMind  
dvijothamcs@gmail.com

**Huan Zhang†**  
UIUC  
huan@huan-zhang.com

## Abstract

Most work on the formal verification of neural networks has focused on bounding the set of outputs that correspond to a given set of inputs (for example, bounded perturbations of a nominal input). However, many use cases of neural network verification require solving the inverse problem, or over-approximating the set of inputs that lead to certain outputs. We present the INVPROP algorithm for verifying properties over the preimage of a linearly constrained output set, which can be combined with branch-and-bound to increase precision. Contrary to other approaches, our efficient algorithm is GPU-accelerated and does not require a linear programming solver. We demonstrate our algorithm for identifying safe control regions for a dynamical system via backward reachability analysis, verifying adversarial robustness, and detecting out-of-distribution inputs to a neural network. Our results show that in certain settings, we find over-approximations over  $2500\times$  tighter than prior work while being  $2.5\times$  faster. By strengthening robustness verification with output constraints, we consistently verify more properties than the previous state-of-the-art on multiple benchmarks, including a large model with 167k neurons in VNN-COMP 2023. Our algorithm has been incorporated into the  $\alpha,\beta$ -CROWN verifier, available at <https://abcrown.org>.

## 1 Introduction

Applying neural networks to safety-critical settings often requires reasoning about constraints on the inputs and outputs of a neural network. For example, for a physical system controlled by a neural network policy, it is of interest to understand which initial states will lead to an unsafe state such as colliding with an obstacle. Formal verification of neural networks seeks to provide provable guarantees demonstrating that networks satisfy formal specifications on their inputs and outputs. Most work to date has focused on developing algorithms that can bound the outputs of a neural network given constraints on the inputs, which can be used for applications such as analyzing the robustness of a neural network to perturbations of a given input [Wong and Kolter, 2018, Dvijotham et al., 2018, Zhang et al., 2018a, Raghunathan et al., 2018, Gehr et al., 2018].

In this work, we address the inverse problem of over-approximating a neural network’s preimage: given a set of outputs  $\mathcal{S}_{\text{out}}$  (described by linear constraints on the network output), we seek to find a set that provably contains all inputs that lead to such outputs. For example, for our neural network policy, this would correspond to the states that collide with an obstacle one step in the future. Though the verification problem is already challenging due to non-convexity and high dimensionality, this new problem is even more difficult since neural networks are generally not invertible.

---

\* Equal contribution, † Equal advising

Instructions for reproducing our results are available at <https://github.com/kothasuhas/verify-input>  
37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Specifically, representative efficient verifiers (such as state-of-the-art bound-propagation-based methods [Zhang et al., 2022]) can only compute bounds utilizing constraints on the input and critically depend on having tight bounds on intermediate activations. In the setting of this paper, however, the bounds derived from the input constraints are almost vacuous, since the only constraints on the input are that it should be from the valid input domain. We efficiently solve this problem by significantly generalizing the existing bound propagation-based verification framework, allowing one to leverage output constraints when tightening the intermediate activations. Our contributions are as follows:

- We develop an effective bound propagation framework, Inverse Propagation for Neural Network Verification (INVPROP), for the *inverse verification problem* for neural networks, i.e., the problem of over-approximating the set of inputs that leads to a given set of outputs. Importantly, INVPROP requires no linear programming solver and can compute bounds on any intermediate layer.
- We unify INVPROP and traditional bound propagation into a more general verification framework, allowing us to connect our method to standard tools, such as the state-of-the-art bound propagation tool  $\alpha,\beta$ -CROWN [Zhang et al., 2018a, Xu et al., 2021, Wang et al., 2021, Zhang et al., 2022, Brix et al., 2023, Müller et al., 2023]. Our contribution allows  $\alpha,\beta$ -CROWN to tighten intermediate bounds with respect to output constraints, which could not be done by the original tool.
- We demonstrate that tight inverse verification requires multiple iterative refinements of intermediate bounds. While layer bounds in standard bound propagation only depend on the bounds of their predecessors, INVPROP incorporates bounds of *all* layers in the network.
- We improve the state of the art on a control benchmark [Rober et al., 2022a,b] by providing  $2500\times$  tighter bounds,  $2.5\times$  faster, for a Double Integrator and  $257\times$  tighter bounds,  $3.29\times$  faster, for a 6D Quadrotor. Furthermore, we demonstrate that INVPROP can strengthen robustness verification with output constraints and verify more robustness properties in less time compared to existing tools. Finally, we demonstrate its applicability in OOD detection.

## 2 Background and Problem Setup

### 2.1 Notation

We use  $[L]$  for  $L \in \mathbb{N}$  to refer to the set  $\{1, 2, \dots, L\}$ ,  $\mathbf{W}_{:,j}^{(i)}$  to refer to column  $j$  of the matrix  $\mathbf{W}^{(i)}$ ,  $[\cdot]_+$  to refer to  $\max(0, \cdot)$ , and  $[\cdot]_-$  to refer to  $-\min(0, \cdot)$ . We use boldface symbols for vectors and matrices (such as  $\mathbf{x}^{(i)}$  and  $\mathbf{W}^{(i)}$ ) and regular symbols for scalars (such as  $x_j^{(i)}$ ). We use  $\mathbf{x} \odot \mathbf{y}$  to denote element-wise multiplication of vectors  $\mathbf{x}, \mathbf{y}$ .

We define an  $L$  layer ReLU neural network by its weight matrices  $\mathbf{W}^{(i)}$  and bias vectors  $\mathbf{b}^{(i)}$  for  $i \in [L]$ . The output of the neural network for the input  $\hat{\mathbf{x}}^{(0)}$  from a bounded input domain  $\mathcal{X}$  is computed by alternately applying linear layers  $\mathbf{x}^{(i)} = \mathbf{W}^{(i)}\hat{\mathbf{x}}^{(i-1)} + \mathbf{b}^{(i)}$  and ReLU layers  $\hat{\mathbf{x}}^{(i)} = \max(\mathbf{0}, \mathbf{x}^{(i)})$  until we receive the output  $\mathbf{x}^{(L)}$  (which we refer to as the logits). Note that we treat softmax as a component of the loss function, not the neural network.

### 2.2 Problem Statement

Given a neural network  $f : \mathcal{X} \subseteq \mathbb{R}^{\text{in}} \rightarrow \mathbb{R}^{\text{out}}$  and an output constraint  $\mathcal{S}_{\text{out}} \subseteq \mathbb{R}^{\text{out}}$ , we want to compute  $f^{-1}(\mathcal{S}_{\text{out}}) \subseteq \mathcal{X}$ . Since precisely computing or expressing  $f^{-1}(\mathcal{S}_{\text{out}})$  is an intractable problem in general, we strive to compute a tight over-approximation  $\mathcal{S}_{\text{over}}$  such that  $f^{-1}(\mathcal{S}_{\text{out}}) \subseteq \mathcal{S}_{\text{over}}$ . In particular, we target the convex hull of the preimage via a cutting-plane representation.  $\mathcal{S}_{\text{out}}$  will be defined by a set of linear constraints parameterized by  $\mathbf{H}f(\mathbf{x}) + \mathbf{d} \leq \mathbf{0}$  in this work.

### 2.3 Applications

**Backward Reachability Analysis for Neural Feedback Loops.** Establishing safety guarantees for neural network policies is a challenging task. One problem of interest is to find a set of initial states that does not reach a particular set of future states under the neural network policy. This can be helpful in collision avoidance or control with safety constraints. For example, consider a discrete-time double integrator controller [Hu et al., 2020] where the state at time  $t + 1$  can be directly computed based on state at time  $t$  following the equation

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \mathbf{x}_t + \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \pi(\mathbf{x}_t)$$

with policy  $\pi : \mathbb{R}^2 \rightarrow \mathbb{R}$ . If there is an obstacle in the room covering the region  $[4.5, 5.0] \times [-0.25, 0.25]$ , it is of interest to understand which states will enter the unsafe region in the next time-step. We can represent this obstacle set with linear constraints that define  $\mathcal{S}_{\text{out}}$ .

$f^{-1}(\mathcal{S}_{\text{out}})$  denotes the set of states  $\mathbf{x}_t$  such that  $\mathbf{x}_{t+1}$  lies in the unsafe region given the control policy  $\pi$ . Overapproximating this set allows us to define the set of states to avoid one timestep in advance. We can compose  $f^{-1}$  with itself  $t$  times to obtain the set of initial states where  $\pi$  would drive the system to the unsafe set after  $t$  steps.

**Robustness Verification** Adversarial robustness is one classic problem for neural network verification where verifiers assess whether all perturbations of an input yield the same classification. INVPROP can be used to speed up this verification by exploiting an implicit output constraint. We follow Wang et al. [2021] and use the canonical form of verifying a property by proving that  $\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) > 0$ .  $f(\mathbf{x})$  will be negative if and only if  $\mathbf{x}$  is an adversarial example. This minimization problem can be rewritten as  $\inf_{\mathbf{x} \in \mathcal{X} \wedge f(\mathbf{x}) \leq 0} f(\mathbf{x})$  [Yang et al., 2021b]. If no adversarial example exists, the infimum is computed over an empty set and returns positive infinity, indicating local robustness. If at least one adversarial example exists, the result is guaranteed to be negative. With the output constraint  $f(\mathbf{x}) \geq 0$ , we only analyze inputs that yield an incorrect classification, reducing the search space and tightening verification bounds.

**Determining what inputs lead to confident predictions.** It is challenging to train a classifier that knows when it doesn’t know. A simple and effective approach is to train with an additional logit representing OOD abstention. Labeled data for this additional class is generated via outlier exposure, or adding synthetic training data known to be OOD [Hendrycks et al., 2018, Chen et al., 2021].

Consider the logits  $y$  produced by a binary classifier trained in this manner. The classifier can classify in-distribution data by comparing  $y_0$  and  $y_1$ , the logits corresponding to the two in-distribution labels. The quantity  $\max(y_0, y_1) - y_2$ , known as the logit gap, can be used to identify out-of-distribution data (Fig. 7 in Appendix A). Can the logit gap correctly identify data far from the training data as being OOD? To answer this question, we can compute the preimage of the set

$$\mathcal{S}_{\text{out}} = \{\mathbf{y} : \max(y_0, y_1) \geq y_2\}$$

We detail how to model this set with linear constraints in Appendix F.1. INVPROP enables us to over-approximate  $f^{-1}(\mathcal{S}_{\text{out}})$ , answering whether the classifier learned to correctly identify OOD data.

### 3 Approach

#### 3.1 Convex over-approximation

Suppose we want to find a convex over-approximation of  $f^{-1}(\mathcal{S}_{\text{out}})$ . The tightest such set is its convex hull, which is the intersection of all half-spaces that contain  $f^{-1}(\mathcal{S}_{\text{out}})$  [Boyd et al., 2004]. This intersection is equivalent to  $\bigcap_{\mathbf{c} \in \mathbb{R}^n} \{\mathbf{x} : \mathbf{c}^\top \mathbf{x} \geq \min_{\mathbf{x}' \in f^{-1}(\mathcal{S}_{\text{out}})} \mathbf{c}^\top \mathbf{x}'\}$ , which means that we can build an over-approximation by taking this intersection for finitely many  $\mathbf{c}$ . Furthermore, replacing the minimization problem with a lower bound to its value still yields a valid half-space, where a tighter bound yields a tighter approximation.

We focus on convex over-approximations for two reasons: First, checking that the preimage satisfies a linear constraint  $\mathbf{c}^\top \mathbf{x} + d \geq 0$  is equivalent to minimizing the linear function  $\mathbf{x} \mapsto \mathbf{c}^\top \mathbf{x}$  over the preimage, which is in turn equivalent to minimizing the linear function over  $\text{CONV}(f^{-1}(\mathcal{S}_{\text{out}}))$  [Boyd et al., 2004]. Second, the convex hull and its tractable over-approximations are conveniently represented as intersections of linear constraints on the preimage, each of which can be quickly computed after a single precomputation phase to tighten bounds on intermediate neurons using the INVPROP algorithm outlined in the next section.

For the above reasons, we will focus on solving the following constrained optimization problem.

$$\min_{\mathbf{x}} \quad \mathbf{c}^\top \mathbf{x}; \quad \text{s.t. } \mathbf{x} \in \mathcal{X}; \quad f(\mathbf{x}) \in \mathcal{S}_{\text{out}} \quad (1)$$

Note that this differs from the widely studied forward verification problem, which can be phrased as

$$\min_{\mathbf{x} \in \mathcal{S}_{\text{in}}} \quad \mathbf{c}^\top f(\mathbf{x})$$

where  $\mathcal{S}_{\text{in}}$  is a set representing constraints on the input, and the goal of the verification is to establish that the output  $f(\mathbf{x})$  satisfies a linear constraint of the form  $\mathbf{c}^\top f(\mathbf{x}) \geq d$  for all  $\mathbf{x} \in \mathcal{S}_{\text{in}}$ .

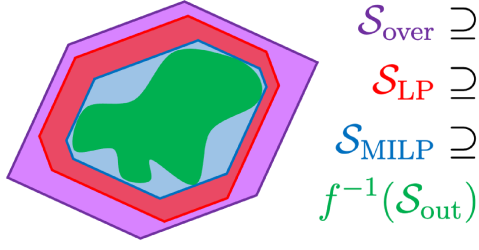


Figure 1: **Visualization of relaxations.** The inner green region depicts the true  $f^{-1}(\mathcal{S}_{\text{out}})$ , the blue relaxation depicts the intersection of finite half-spaces solved via MILP, the red relaxation displays the same via LP, and the purple relaxation displays the same via INVPROP. Though this diagram displays looseness, we provide a comprehensive methodology to reduce the error in all three relaxations up to arbitrary precision (Section 3.3).

### 3.2 The INVPROP Algorithm

As a brief overview of our method, we first construct the Mixed Integer Linear Program (MILP) for solving this optimization problem, which generates the over-approximation  $\mathcal{S}_{\text{MILP}}$  when solved for finitely many  $c$ . Then, we relax the MILP to a Linear Program (LP), which can construct the over-approximation  $\mathcal{S}_{\text{LP}}$ . Finally, we relax the LP via its Lagrangian dual, which will be used to construct our over-approximation  $\mathcal{S}_{\text{over}}$ . This chain of relaxations is visualized in Figure 1.

**The Mixed Integer Linear Programming (MILP) Formulation** For feed-forward ReLU neural networks, the non-linearities from the max operator can be encoded via integer variables, and this problem admits a MILP encoding similar to prior work in adversarial robustness [Tjeng et al., 2017]. Problem (1) is equivalent to:

$$\min_{\mathbf{x}, \hat{\mathbf{x}}} \quad \mathbf{c}^\top \mathbf{x} \quad (2a)$$

$$\text{s.t.} \quad \mathbf{x} \in \mathcal{X}; \quad \hat{\mathbf{x}}^{(0)} = \mathbf{x}; \quad \mathbf{H}\mathbf{x}^{(L)} + \mathbf{d} \leq \mathbf{0}; \quad (2b)$$

$$\mathbf{x}^{(i)} = \mathbf{W}^{(i)}\hat{\mathbf{x}}^{(i-1)} + \mathbf{b}^{(i)} \quad i \in [L]; \quad (2c)$$

$$\hat{\mathbf{x}}^{(i)} = \max(0, \mathbf{x}^{(i)}); \quad i \in [L-1] \quad (2d)$$

**The Linear Programming (LP) Formulation** Unfortunately, finding an exact solution to this MILP is NP-complete [Katz et al., 2017]. To sidestep the intractability of exact verification, we can compute lower bounds for this program via its convex relaxation. We consider bounds on the outputs of intermediate layers:

$$l_j^{(i)} \leq x_j^{(i)} \leq u_j^{(i)}$$

Based on these bounds, we can take the ReLU triangle relaxation [Ehlers, 2017] to get the LP

$$\min_{\mathbf{x}, \hat{\mathbf{x}}} \quad \mathbf{c}^\top \mathbf{x} \quad (3a)$$

$$\text{s.t.} \quad \mathbf{x} \in \mathcal{X}; \quad \mathbf{l}^{(0)} \leq \mathbf{x} \leq \mathbf{u}^{(0)}; \quad \hat{\mathbf{x}}^{(0)} = \mathbf{x}; \quad \mathbf{H}\mathbf{x}^{(L)} + \mathbf{d} \leq \mathbf{0} \quad (3b)$$

$$\mathbf{x}^{(i)} = \mathbf{W}^{(i)}\hat{\mathbf{x}}^{(i-1)} + \mathbf{b}^{(i)} \quad (3c)$$

$$\mathbf{0} \leq \hat{\mathbf{x}}^{(i)}; \quad \mathbf{x}^{(i)} \leq \hat{\mathbf{x}}^{(i)}; \quad \hat{\mathbf{x}}^{(i)} \leq \frac{\mathbf{u}^{(i)}}{\mathbf{u}^{(i)} - \mathbf{l}^{(i)}} \odot (\mathbf{x}^{(i)} - \mathbf{l}^{(i)}); \quad i \in [L] \quad (3d)$$

where the bounds  $\mathbf{l}^{(0)} \leq \mathbf{x} \leq \mathbf{u}^{(0)}$  are either the bounds implicit in  $\mathcal{X}$ , or a refinement of these obtained via previous rounds of bound propagation or input branching (see Algorithm 1 for details). Most efficient neural network verifiers do not solve an LP formulation of verification directly because LP solvers are often slow for problem instances involving neural networks. Instead, the bound propagation framework [Zhang et al., 2018a, Wang et al., 2021, Zhang et al., 2022] is a practical and efficient way to lower bound the LP relaxation of the forward verification problem. However, there are *two major roadblocks* to applying existing methods here: Typical bound-propagation cannot directly handle the output constraints (Eq. 3b) and the objective involving input layer variables (Eq. 3a). This is true for optimizing bounds on the input, intermediate layers, and the output layer, all of which need to be iteratively tightened as demonstrated later.

**The Inverse Propagation (INVPROP) Formulation** By changing the LP above to optimize the quantity  $\hat{x}_j^{(0)}$  (input layer) or  $x_j^{(i)}$  (intermediate layers) for  $i \in [L - 1]$ , the bounds for the  $j$ -th neuron of layer  $i$  can be tightened in separate LP calls. However, this program is too expensive to be run multiple times for each neuron.

Inspired by the success of CROWN-family neural network verifiers [Zhang et al., 2018a, Xu et al., 2021, Wang et al., 2021, Zhang et al., 2022], we efficiently lower bound the solution of the LP by optimizing its Lagrangian dual. This dual is highly structured [Wong and Kolter, 2018], allowing us to bound input half-spaces  $c^\top x$  and intermediate bounds  $l_j^{(i)}, u_j^{(i)}$  by closed-form expressions of the dual variables. Our main generalization is the ability to optimize input or intermediate layer bounds with the output constraints **after** them in the neural network, as shown in the following theorem.

**Theorem 1** (Bounding input half-spaces). *Given an output set  $\mathcal{S}_{out} = \{y : Hy + d \leq 0\}$  and vector  $c$ ,  $g_c(\alpha, \gamma)$  is a lower bound to the linear program in (3) for  $0 \leq \alpha \leq 1$ ,  $\gamma \geq 0$ , and  $g_c$  defined via*

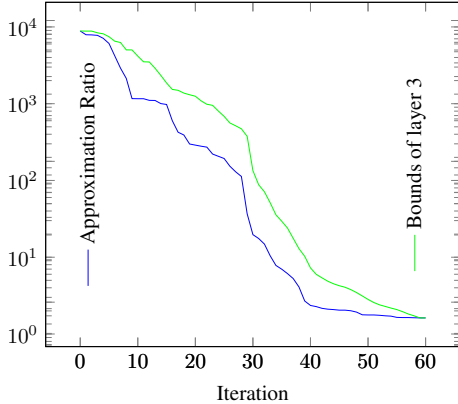
$$g_c(\alpha, \gamma) = \left[ c^\top - \nu^{(1)\top} W^{(1)} \right]_+ l^{(0)} - \left[ c^\top - \nu^{(1)\top} W^{(1)} \right]_- u^{(0)} \\ - \sum_{i=1}^L \nu^{(i)\top} b^{(i)} + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^\pm(i)} \left[ \frac{u_j^{(i)} l_j^{(i)} [\hat{\nu}_j^{(i)}]_+}{u_i^{(j)} - l_i^{(j)}} \right]$$

where every term can be directly recursively computed via

$$\mathcal{I}^{-(i)} = \{j : u_j^{(i)} \leq 0\}; \quad \mathcal{I}^{+(i)} = \{j : l_j^{(i)} \geq 0\}; \quad \mathcal{I}^{\pm(i)} = \{j : l_j^{(i)} < 0 < u_j^{(i)}\} \\ \nu^{(L)} = -\gamma; \quad \hat{\nu}_j^{(i)} = \nu^{(i+1)\top} W_{:,j}^{(i+1)} \quad \forall i \in [L - 1] \\ \nu_j^{(i)} = \begin{cases} \hat{\nu}_j^{(i)}, & j \in \mathcal{I}^{+(i)} \\ 0, & j \in \mathcal{I}^{-(i)} \\ \frac{u_j^{(i)}}{u_i^{(j)} - l_i^{(j)}} [\hat{\nu}_j^{(i)}]_+ - \alpha_j^{(i)} [\hat{\nu}_j^{(i)}]_-, & j \in \mathcal{I}^{\pm(i)} \end{cases} \quad \forall i \in [L - 1]$$

*Proof.* Full proof is presented in Appendix D. □

In Appendix C, we show how to bound intermediate neurons in the network using a similar approach. Since the intermediate bounds might depend on bounds on neurons in subsequent layers (due to the output constraint), we cannot simply optimize bounds in a single forward pass layer by layer, unlike prior work such as  $\alpha$ -CROWN [Xu et al., 2021]. Instead, we must iteratively tighten intermediate layer bounds with respect to the tightest bounds on all neurons computed thus far. This iterative approach can tighten the initially loose bounds by several orders of magnitude, as shown in Figure 2. After performing this procedure once, the intermediate bounds can be used to tightly lower bound  $c^\top x$  for any  $c$  via Theorem 1. Therefore, this computation can be shared across all the constraints  $c$  we use to describe  $\mathcal{S}_{over}$ . Our algorithm can be expressed in terms of forward/backward passes through layers of the neural network and implemented via standard deep learning modules in libraries like PyTorch [Paszke et al., 2019]. Since all of the operations are auto-differentiable, we can tighten our lower bound using standard gradient ascent (projected by the dual variable constraints).



**Figure 2: Necessity of Iterative Tightening.** Our approach enables us to iteratively tighten the bounds of all layers, with each iteration allowing for a smaller approximation ratio with respect to the true preimage. Green: Sum of bound intervals for all neurons in the third layer (second hidden layer); Blue: ratio between volumes of over-approximation and preimage. Measured for step  $t = 10$  of the control benchmark defined in Section 4.1. Note that the improvement from iterative tightening is two orders of magnitude for intermediate bounds, and four orders of magnitude for the volume of the over-approximation.

---

**Algorithm 1** INVPROP. Can be applied to branches for non-convex overapproximations. Lower and upper bound of all neurons and all constraints are optimized using distinct  $\alpha$  and  $\gamma$  values in  $g$ .

---

```

Initialize  $l^{(i)}, u^{(i)}$  via cheap bound propagation methods for the forward verification problem
while lower bounds for  $c^\top \hat{x}^{(0)}$  are improving do
  for  $i \in \{L-1, L-2, \dots, 1, 0\}$  do
    for  $j \in \text{layer } i \text{ neurons}, b \in \{\text{lower, upper}\}$  do
      Optimize  $g_{ij}^b$  from Theorem 2 via gradient ascent to improve bound on neuron  $j$  in layer  $i$ 
      with sense  $b$  (lower/upper)
      if  $b = \text{upper}$  then update  $u_j^{(i)}$  else update  $l_j^{(i)}$ 
    Optimize  $g_c$  from Theorem 2 via gradient ascent to improve lower bound on  $c^\top x$  for all  $c$ 

```

---

**Connection to forward verification** Our bound in Theorem 1 introduces the dual variable  $\gamma$ , which enforces the output constraint during optimization. In fact, we can use this variable to get a better conceptual interpretation of our result. For the optimization problem in (1), taking the dual with respect to the constraint  $\mathbf{H}f(x) + \mathbf{d} \leq \mathbf{0}$  yields the lower bound

$$\begin{aligned} \max_{\gamma} \min_x \quad & c^\top x + \gamma^\top (\mathbf{H}f(x) + \mathbf{d}) \\ \text{s.t.} \quad & x \in \mathcal{X}; \quad \gamma \geq 0 \end{aligned}$$

The objective can be represented as minimizing a linear function of the output of a residual neural network with a skip connection from the input to the output layer, subject to constraints on the input. Now,  $f(x)$  appears in the objective, similar to the standard forward verification problem with an augmented network architecture. Similarly, optimizing an intermediate bound is equivalent to a skip connection from layer  $i$  to the output (Appendix E). These connections allow us to implement our method using standard verification tools [Brix et al., 2023, Xu et al., 2021, 2020].

When  $f$  is a feedforward ReLU network, the lower bound described in this section is precisely the same as Theorem 1 (since both are solving the dual of the same linear program). This shows that the introduction of  $\gamma$  constitutes our generalization to the bound propagation framework.

**Selection of  $c$**  The dependence of the optimization on  $c$  is not a major limitation. The bounds of all intermediate neurons are optimized only with respect to the input and output constraints, not the hyperplanes used to describe the input half-space. Thus, their optimized bounds can be shared across the optimization of all input hyperplanes. For applications such as adversarial robustness, the box constraint of two hyperplanes per dimension could be chosen to scalably bound the input. The selection of appropriate  $c$  depends on the application and is not our main focus.

### 3.3 Branch and Bound

Though more rounds of iterative tightening will lower the gap in  $\mathcal{S}_{\text{over}} \supseteq \mathcal{S}_{\text{LP}}$ , our current formulation still faces two sources of looseness: the gap in  $\mathcal{S}_{\text{LP}} \supseteq \mathcal{S}_{\text{MILP}}$  and the gap in  $\mathcal{S}_{\text{MILP}} \supseteq f^{-1}(\mathcal{S}_{\text{out}})$ . To overcome both of these issues, we can make use of branching [Bunel et al., 2020]. While there are several possibilities here, we focus on input branching, which gave the biggest empirical gains in our experiments described in Section 4. More concretely, we divide the input domain  $\mathcal{X} = [l^{(0)}, u^{(0)}]$  into several regions by choosing a coordinate  $i$  and computing

$$\mathcal{X}_a = \mathcal{X} \cap \{x : x_i \geq s_i\}, \mathcal{X}_b = \mathcal{X} \cap \{x : x_i \leq s_i\}$$

so that  $\mathcal{X} = \mathcal{X}_a \cup \mathcal{X}_b$ . Doing this recursively, we obtain several regions and can compute an overapproximation of  $f^{-1}(\mathcal{S}_{\text{out}})$  when the input is in each of those regions, and take a union of the resulting sets. The finer the input domain of our over-approximation, the tighter the approximation.

## 4 Results

We evaluate INVPROP on three different benchmarks. Across benchmarks, we find orders of magnitude improvement over prior work, and our methods work even for relatively large networks (167k neurons) and high dimensionality inputs (8112 dimensions). All implementation details are described in Appendix F and the utilized hardware is described in Appendix G.

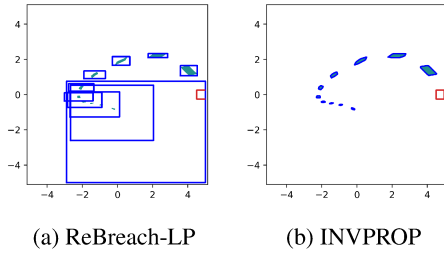


Figure 3: **Double Integrator**. Each green region represents the preimage of the system one to ten steps back from the obstacle (red box). Each preimage is approximated 1 million samples. The blue bounding boxes represent over-approximations, and INVPROP tightly over-approximates preimages even ten steps back.

Table 1: **Double Integrator Results**. INVPROP vs SOTA (from Figure 3). We are significantly tighter and faster than ReBreach-LP method even without input branching.

| METHOD          | REBREACH-LP  | INVPROP      |
|-----------------|--------------|--------------|
| INPUT BRANCHING | YES          | NO           |
| APPROX RATIO ↓  | 4021.47      | 1.46         |
| TIME (SEC)      | 42.86 ± 0.04 | 17.89 ± 0.03 |

#### 4.1 Backward Reachability Analysis for Neural Feedback Loops

**Double Integrator** After combining the state-feedback control policy  $\pi$  of the benchmark with the state transition function from Section 2.3, (see Section F.2), we get a three layer MLP with 12, 7, and 2 neurons, which is typical for applications in this setting. We model multiple time steps via composing copies of this function.<sup>1</sup> Moreover, we leverage intermediate bounds for time step  $t$  to bound the  $t + 1$  time step transition. We measure the tightness of an over-approximation using its Approximation Ratio, defined as  $\frac{\text{vol}(S_{\text{over}})}{\text{vol}(f^{-1}(S_{\text{out}}))}$ . Both of these volumes are heuristically estimated with 1 million samples of the input space. While INVPROP allows the optimization of arbitrary input cutting planes, we use 40 planes with slopes of equally distributed angles.

We find that INVPROP is significantly tighter and faster than ReBreach-LP, the SOTA available method for this problem [Rober et al., 2022a,b].<sup>2</sup> As evident in Figure 3, ReBreach-LP suffers from increasingly weak bounds as  $t$  increases, whereas our approach is able to compute tight bounds for all  $t$ . Different to their implementation, we are able to leverage the output constraint to improve the intermediate neuron bounds. Furthermore, we can iteratively tighten these bounds with respect to each other. Even though we optimize many more quantities, our efficient bound propagation allows us to solve the problem faster overall. We provide a visualization of the improvement of the approximation ratio score and intermediate bounds over time in Figure 2. Both our increased tightness and speed are quantified in Table 1.

**6D Quadrotor** We consider over-approximating the preimage of the linearized quadrotor discussed in Figure 15 of Rober et al. [2022b]. This quadrotor has a 6-dimensional state space with 3 dimensions dedicated to position and is given by the dynamics

$$\mathbf{x}_{t+1} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}_t + \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \pi(\mathbf{x}_t)$$

The policy, after performing the same encoding as the double integrator and clipping policy outputs, forms a five layer MLP with 26, 26, 9, 9, and 3 neurons. For these dynamics, we want to prove that the initial state range of  $[-5.25, -4.75] \times [-.25, .25] \times [2.25, 2.75] \times [0.95, 0.99] \times [-0.01, 0.01] \times [-0.01, 0.01]$  (black box in Figure 4) will never collide with the obstacle at  $[-1, 1] \times [-1, 1] \times [1.5, 3.5] \times [-1, 1] \times [-1, 1] \times [-1, 1]$  (red box). We try to find the tightest bounding box in the 3 position dimensions, reflecting a total of 6 input cutting planes (one for each direction in each dimension). To benchmark against prior work, we run ReBreach-LP with no output partitioning and

<sup>1</sup>For example, the 10 time step transition function can be represented as a 21 layer MLP (after fusing consecutive linear layers in the composition).

<sup>2</sup>We do not compare with Everett et al. [2022] which improves upon ReBreach-LP, as their implementation faces numerical instability when applied for ten timesteps. However, preliminary experiments show that we also outperform their bounds.

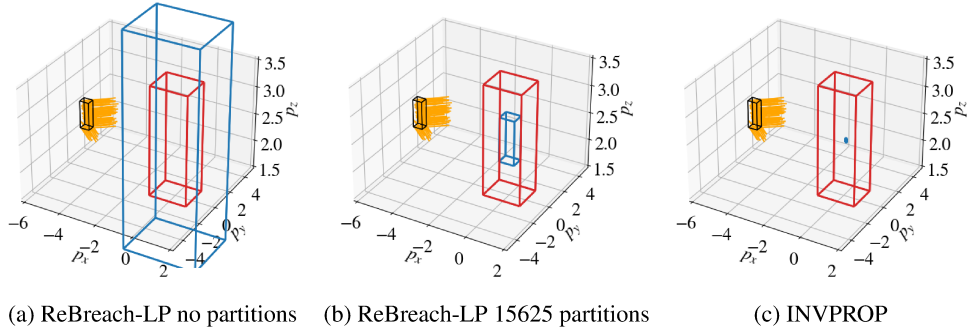


Figure 4: **6D Quadrotor.** We visualize the three dimensions of the 6D quadrotor that specify physical position. We consider the initial state range of the black box and simulate a few one-step trajectories, indicated by the orange lines. Our obstacle is the red box. For the three images, the blue box represents the over-approximation computed by (a) ReBreach-LP with no partitioning, (b) ReBreach-LP with 15625 partitions, and (c) INVPROP, respectively. A tighter blue box is better.

Table 2: **Quadrotor Results.** We compare the tightness of the over-approximations and run time (mean and std taken over 5 runs).

| METHOD       | REBREACH-LP    | REBREACH-LP      | INVPROP          |
|--------------|----------------|------------------|------------------|
| PARTITIONS   | N/A            | 15625            | N/A              |
| APPROX VOL ↓ | 64             | 0.064            | 0.0000249        |
| TIME (SEC)   | $11.8 \pm 2.2$ | $662.2 \pm 31.4$ | $213.8 \pm 11.1$ |

ReBreach-LP with 15625 output partitions. We run our algorithm with no branching. In Figure 4, we compare these three strategies, and we find that our over-approximation of the pre-image (in blue) is smaller than the state-of-the-art over-approximations by 257 times while being 3.29 times faster. Therefore, our algorithm scales well to higher dimensional control examples.

## 4.2 Robustness Verification

Similar to us, Yang et al. [2021b] encode the implicit output constraint of the local robustness verification query to tighten the bounds on neurons. However, they need an LP solver, which does not scale to large problems. The state-of-the-art verification toolkit  $\alpha, \beta$ -CROWN [Zhang et al., 2018a, Xu et al., 2021, Wang et al., 2021, Zhang et al., 2022] is able to scale to large networks, but does not currently utilize the implicit output constraint. We demonstrate the benefit of encoding the output constraint by extending  $\alpha, \beta$ -CROWN and comparing the performance on the benchmark used by Yang et al. [2021b] as well as benchmarks from the VNN-COMP 2023 [Brix et al., 2023].

**MNIST** Yang et al. [2021b] provides four networks with ReLU activation functions and dense linear layers of 2/6/6/9 layers of 100/100/200/200 neurons, each trained on MNIST. For each network, 50 inputs were tested for local robustness. For the complete benchmark definition, we refer to Yang et al. [2021b]. The DeepSRGR results are taken from Yang et al. [2021b], they do not report a timeout for their experiments. Both  $\alpha, \beta$ -CROWN and  $\alpha, \beta$ -CROWN+INVPROP were run with a per-input timeout of 5 minutes. Except for the input bounds, all bounds of all layers of each network are tightened using output constraints. We report the results in Table 3. Notably, we can verify more instances than the SOTA tool  $\alpha, \beta$ -CROWN, in sometimes less than one fifth the average runtime. We include a more detailed comparison between the results for  $\alpha, \beta$ -CROWN and  $\alpha, \beta$ -CROWN+INVPROP in Appendix H.

**VNN-COMP '23: YOLO** In 2023, the VNN-COMP contained YOLO as an unscored benchmark for object detection. It is a modified version of YOLOv2 [Redmon and Farhadi, 2016] with a network of 167k neurons and 5 residual layers, and is available at [Zhong, 2023, Brix, 2023]. The network processes  $3 \times 52 \times 52$  images and uses convolutions, average pooling and ReLU activations. This benchmark is well suited for INVPROP, as the definition of the adversarial examples is a conjunction of constraints over the output. Therefore, a strong output constraint can be used to tighten the bounds of intermediate layers. The benchmark consists of 464 instances, of which 72 were randomly selected for the VNN-COMP. For our comparison, we remove those 348 instances that  $\alpha, \beta$ -CROWN verifies as robust without tightening the bounds of any intermediate layer. Those instances are verified within

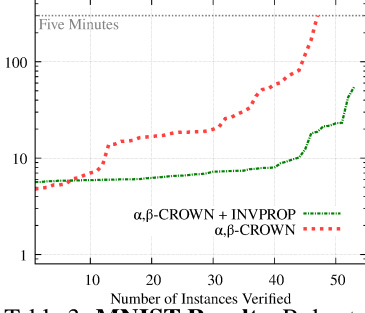


Figure 5: **YOLO Results.** Runtime comparison of  $\alpha,\beta$ -CROWN and  $\alpha,\beta$ -CROWN+INVPROP on the YOLO benchmark (167k neurons and 5 residual layers). For the comparison, only those test instances were used that could not immediately be verified by  $\alpha,\beta$ -CROWN without any iterative tightening of intermediate layer bounds.  $\alpha,\beta$ -CROWN+INVPROP can verify more properties and is faster for almost all instances.

Table 3: **MNIST Results.** Robustness verification with DeepSRGR (output constraints via LP solver),  $\alpha,\beta$ -CROWN (GPU support) and  $\alpha,\beta$ -CROWN+INVPROP (output constraints with GPU support)

| NETWORK      | DEEPSRGR |               | $\alpha,\beta$ -CROWN |               | $\alpha,\beta$ -CROWN + INVPROP |               |
|--------------|----------|---------------|-----------------------|---------------|---------------------------------|---------------|
|              | VERIF.   | AVG. TIME [S] | VERIF.                | AVG. TIME [S] | VERIF.                          | AVG. TIME [S] |
| FFN4 (3x100) | 35       | 781           | 45                    | 7.4           | 45                              | 6.4           |
| FFN5 (6x100) | 31       | 1689          | 38                    | 6.7           | 39                              | 8.7           |
| FFN6 (6x200) | 31       | 6178          | 29                    | 5.4           | 30                              | 1.0           |
| FFN7 (9x200) | 36       | 8960          | 37                    | 3.2           | 40                              | 1.9           |

less than four seconds each by both methods, with no room for improvement. We compare the performance of  $\alpha,\beta$ -CROWN and  $\alpha,\beta$ -CROWN + INVPROP on the remaining 116 instances in Figure 5.  $\alpha,\beta$ -CROWN can verify 48 instances, all other instances reach the timeout of five minutes. After extending  $\alpha,\beta$ -CROWN with INVPROP on the last two intermediate layers, almost all instances can be solved faster, and 6 previously timed out instances become verifiable.

### 4.3 OOD Detection

Consider the calibrated OOD detector presented as discussed in Section 2.3, encoded by a four layer MLP with 200, 200, 3, and 2 neurons. We over-approximate the set of inputs which induce a sufficiently high in-distribution (ID) confidence (measured by  $\max\{y_0, y_1\} > y_2$ ) using 40 hyperplanes of equal slope, pictured in green in Figure 6. This set is non-convex, making the convex hull a poor over-approximation. With 4 input space branches, we get a much tighter over-approximation, as shown in the right plot. We compare the performance of our approach with and without branching over the input space with the MILP baseline (see Table 4). This demonstrates a simple proof-of-concept for how INVPROP can be used for verifying some calibration properties.

## 5 Related Work

**Formally Verified Neural Networks.** There has been a large body of work on the formal verification of neural networks, tackling the problem from the angles of Interval Bound Propagation [Gowal et al., 2018, Gehr et al., 2018], Convex Relaxations [Wong and Kolter, 2018, Shiqi et al., 2018, Salman et al., 2019, Dvijotham et al., 2018], Abstract Interpretation [Singh et al., 2018], LP [Ehlers, 2017], SDP [Raghunathan et al., 2018], SMT [Katz et al., 2017], and MILP [Tjeng et al., 2017]. However, most of this work is for forward verification (i.e., bounding the NN output given a

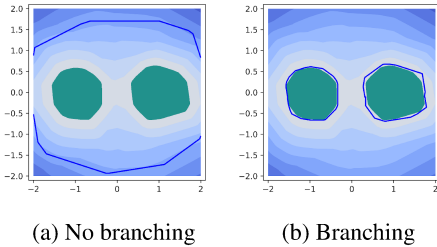


Figure 6: **OOD Detection.** Green: Empirical ID inputs (1 million samples). Blue: border of the verified preimage. Branching is union of preimages for each of 4 quadrants.

Table 4: **OOD Results.** Comparison of methods for over-approximating the preimage of the OOD region (Figure 6). A lower approximation ratio (APPROX RAT.) is better.

| METHOD  | INPUT BRANCH. | APPROX RATIO | TIME (SEC)       |
|---------|---------------|--------------|------------------|
| MILP    | NO            | 1.47         | 1562.26          |
| INVPROP | NO            | 4.39         | $3.77 \pm 0.02$  |
| INVPROP | YES           | 1.14         | $12.02 \pm 0.06$ |

set of inputs). Our work strictly generalizes bound propagation as presented in Xu et al. 2021 since setting  $\gamma = \mathbf{0}$  in Theorem 1 recovers its results.

**Formal Analysis of the Input Domain** Prior work has studied trying to determine the inverse of a neural network [Kindermann and Linden, 1990, Saad and Wunsch, 2007], though these methods are prohibitively slow. Zhang et al. [2018b] used an LP solver to compute changes to a given input that would lead to a change in the classification result while adhering to additional criteria. Yang et al. [2021b] encode the overapproximated network in an LP problem to tighten bounds by incorporating the output constraint. They require an LP solver and we compare our performance against theirs in Section 4.2. Zhong et al. [2023], Wu et al. [2022] similarly tighten intermediate bounds by using an LP solver. Dimitrov et al. [2022] compute input regions that only consist of adversarial examples (i.e., adhere to a given output constraint) that are maximized using an LP solver. Finally, a concurrent work Zhang et al. [2023] targets under- and over-approximating the preimage via LiRPA-based forward verification [Xu et al., 2020] with input-space and ReLU-space partitioning; since our work can be viewed as a generalization of LiPRA, their work may benefit from our results of tightening the intermediate layer bounds using output constraints during bound propagation.

**Formal Verification for Neural Feedback Loops.** Our control application was motivated by the growing body of work on backward reachability analysis and over-approximating states that result in a target set [Everett et al., 2021]. The original method solely utilized input constraints for deriving intermediate bounds. The later development of Everett et al. 2022 improved upon this by optimizing  $\mathbf{l}^{(0)}$  and  $\mathbf{u}^{(0)}$  with respect to output bounds but their implementation faces numerical instability when applied for many iterations. We also support the partitioning over the input space introduced by Rober et al. 2022a. The work of Vincent and Schwager 2021 explores utilizing an LP with complete neuron branching to verify control safety, which can be viewed as a domain-specific implementation of our MILP formulation of the inverse verification problem.

**Certified OOD detection.** There is a wide variety of OOD detection systems [Yang et al., 2021a, Salehi et al., 2022]. Commonly, they are evaluated empirically based on known OOD data [Wang et al., 2022, Liang et al., 2018, Chen et al., 2021]. Therefore, they fail to provide verified guarantees for their effectiveness. In fact, many OOD detection systems are susceptible to adversarial attacks [Schwag et al., 2019, Chen et al., 2022]. Meinke et al. [2021] show how to verify robustness to adversarial examples around given input samples. Berrada et al. [2021] develop a general framework for certifying properties of output distributions of neural networks given constraints on the input distribution. However, this work is still constrained to verifying properties of the outputs given constraints (albeit probabilistic) on the inputs, and INVPROP is able to certify arbitrary regions of the input space that lead to confident predictions.

## 6 Discussion

We present the challenge of over-approximating neural network preimages and provide an efficient algorithm to solve it. By doing so, we demonstrate strong performance on multiple application areas. We believe there is a large scope for future investigation and new applications of INVPROP.

**Limitations** For higher-dimensional instances, our method would best work for box over-approximations ( $2d$  hyperplanes in  $d$  dimensions) and would struggle at more complicated shapes such as spheres. We expect more complex problems will benefit from a domain-specific strategy. Moreover, iteratively refining all of the intermediate bounds utilizing INVPROP faces a quadratic dependence on network depth, as opposed to a linear dependence in traditional forward verification. To mitigate this, output constraints could only be applied to layers close to the output, as we used for robustness. Finally, our branching strategy might not scale to higher dimensions, though this trade-off is well-studied [Bunel et al., 2020, Wang et al., 2021, Palma et al., 2023].

**Potential Negative Social Impact** Our work improves reliable ML through facilitating systems that provably align with practitioner expectations and we expect INVPROP to have positive societal impact. We acknowledge that our improvements may be repurposed as model attacks, though we believe the positive use cases of our technique greatly outweigh current speculation of misuse.

**Acknowledgements** We thank Michael Everett and Nicholas Rober for helpful discussion and feedback on the paper. Huan Zhang acknowledges the support of the Schmidt Futures AI2050 Early Career Fellowship.

## References

- Leonard Berrada, Sumanth Dathathri, Krishnamurthy Dvijotham, Robert Stanforth, Rudy R Bunel, Jonathan Uesato, Sven Gowal, and M Pawan Kumar. Make sure you’re unsure: A framework for verifying probabilistic specifications. *Advances in Neural Information Processing Systems*, 34: 11136–11147, 2021.
- Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.
- Christopher Brix. Vnn-comp 2023 benchmarks. [https://github.com/ChristopherBrix/vnncomp2023\\_benchmarks](https://github.com/ChristopherBrix/vnncomp2023_benchmarks), 2023.
- Christopher Brix, Mark Niklas Müller, Stanley Bak, Taylor T. Johnson, and Changliu Liu. First three years of the international verification of neural networks competition (vnn-comp), 2023. URL <https://arxiv.org/abs/2301.05815>.
- Rudy Bunel, P Mudigonda, Ilker Turkaslan, P Torr, Jingyue Lu, and Pushmeet Kohli. Branch and bound for piecewise linear neural network verification. *Journal of Machine Learning Research*, 21 (2020), 2020.
- Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2021.
- Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Robust out-of-distribution detection for neural networks. In *The AAAI-22 Workshop on Adversarial Machine Learning and Beyond*, 2022. URL [https://openreview.net/forum?id=WMIoz7O\\_DPz](https://openreview.net/forum?id=WMIoz7O_DPz).
- Dimitar Iliev Dimitrov, Gagandeep Singh, Timon Gehr, and Martin Vechev. Provably robust adversarial examples. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=UMfhoMtIaP5>.
- Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy A Mann, and Pushmeet Kohli. A dual approach to scalable verification of deep networks. In *UAI*, volume 1, page 3, 2018.
- Ruediger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, pages 269–286. Springer, 2017.
- Michael Everett, Golnaz Habibi, Chuangchuang Sun, and Jonathan P How. Reachability analysis of neural feedback loops. *IEEE Access*, 9:163938–163953, 2021.
- Michael Everett, Rudy Bunel, and Shayegan Omidshafiei. Drip: Domain refinement iteration with polytopes for backward reachability analysis of neural feedback loops. *arXiv preprint arXiv:2212.04646*, 2022.
- Timon Gehr, Matthew Mirman, Dana Drachler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2018. doi: 10.1109/SP.2018.00058.
- Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.
- Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.
- Haimin Hu, Mahyar Fazlyab, Manfred Morari, and George J. Pappas. Reach-sdp: Reachability analysis of closed-loop systems with neural network controllers via semidefinite programming. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 5929–5934, 2020. doi: 10.1109/CDC42340.2020.9304296.

- Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International conference on computer aided verification*, pages 97–117. Springer, 2017.
- J Kindermann and A Linden. Inversion of neural networks by gradient descent. *Parallel Computing*, 14(3):277–286, 1990. ISSN 0167-8191. doi: [https://doi.org/10.1016/0167-8191\(90\)90081-J](https://doi.org/10.1016/0167-8191(90)90081-J). URL <https://www.sciencedirect.com/science/article/pii/016781919090081J>.
- Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=H1VGkIxRZ>.
- Alexander Meinke, Julian Bitterwolf, and Matthias Hein. Provably robust detection of out-of-distribution data (almost) for free. *arXiv preprint arXiv:2106.04260*, 2021.
- Mark Niklas Müller, Christopher Brix, Stanley Bak, Changliu Liu, and Taylor T. Johnson. The third international verification of neural networks competition (vnn-comp 2022): Summary and results, 2023.
- Alessandro De Palma, Rudy Bunel, Krishnamurthy Dvijotham, M. Pawan Kumar, and Robert Stanforth. Ibp regularization for verified adversarial robustness via branch-and-bound, 2023.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.
- Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016. URL <http://arxiv.org/abs/1612.08242>.
- Nicholas Rober, Michael Everett, Songan Zhang, and Jonathan P How. A hybrid partitioning strategy for backward reachability of neural feedback loops. *arXiv preprint arXiv:2210.07918*, 2022a.
- Nicholas Rober, Sydney M Katz, Chelsea Sidrane, Esen Yel, Michael Everett, Mykel J Kochenderfer, and Jonathan P How. Backward reachability analysis of neural feedback loops: Techniques for linear and nonlinear systems. *arXiv preprint arXiv:2209.14076*, 2022b.
- Emad W. Saad and Donald C. Wunsch. Neural network explanation using inversion. *Neural Networks*, 20(1):78–93, 2007. ISSN 0893-6080. doi: <https://doi.org/10.1016/j.neunet.2006.07.005>. URL <https://www.sciencedirect.com/science/article/pii/S0893608006001730>.
- Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*, 2022. URL <https://openreview.net/forum?id=aRtjVZvbpK>.
- Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. A convex relaxation barrier to tight robustness verification of neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.
- Vikash Sehwal, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, AISec’19*, page 105–116, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368339. doi: 10.1145/3338501.3357372. URL <https://doi.org/10.1145/3338501.3357372>.

- Wang Shiqi, Kexin Pei, Whitehouse Justin, Junfeng Yang, and Suman Jana. Efficient formal safety analysis of neural networks. In *32nd Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2018.
- Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL <https://proceedings.neurips.cc/paper/2018/file/f2f446980d8e971ef3da97af089481c3-Paper.pdf>.
- Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL), jan 2019. doi: 10.1145/3290354. URL <https://doi.org/10.1145/3290354>.
- Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. *arXiv preprint arXiv:1711.07356*, 2017.
- Joseph A. Vincent and Mac Schwager. Reachable polyhedral marching (rpm): A safety verification algorithm for robotic systems with deep neural network components. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9029–9035, 2021. doi: 10.1109/ICRA48506.2021.9561956.
- Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. Beta-CROWN: Efficient bound propagation with per-neuron split constraints for complete and incomplete neural network verification. *Advances in Neural Information Processing Systems*, 34, 2021.
- Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018.
- Haoze Wu, Clark Barrett, Mahmood Sharif, Nina Narodytska, and Gagandeep Singh. Scalable verification of gnn-based job schedulers. *Proc. ACM Program. Lang.*, 6(OOPSLA2), oct 2022. doi: 10.1145/3563325. URL <https://doi.org/10.1145/3563325>.
- Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kaillkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. *Advances in Neural Information Processing Systems*, 33, 2020.
- Kaidi Xu, Huan Zhang, Shiqi Wang, Yihan Wang, Suman Jana, Xue Lin, and Cho-Jui Hsieh. Fast and Complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=nVZtXBI6LNn>.
- Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021a.
- Pengfei Yang, Renjue Li, Jianlin Li, Cheng-Chao Huang, Jingyi Wang, Jun Sun, Bai Xue, and Lijun Zhang. Improving neural network verification through spurious region guided refinement. In Jan Friso Groote and Kim Guldstrand Larsen, editors, *Tools and Algorithms for the Construction and Analysis of Systems*, pages 389–408, Cham, 2021b. Springer International Publishing. ISBN 978-3-030-72016-2.
- Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. *Advances in Neural Information Processing Systems*, 31:4939–4948, 2018a. URL <https://arxiv.org/pdf/1811.00866.pdf>.
- Huan Zhang, Shiqi Wang, Kaidi Xu, Linyi Li, Bo Li, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. General cutting planes for bound-propagation-based neural network verification. *Advances in Neural Information Processing Systems*, 2022.

- Xin Zhang, Armando Solar-Lezama, and Rishabh Singh. Interpreting neural network judgments via minimal, stable, and symbolic corrections. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL [https://proceedings.neurips.cc/paper\\_files/paper/2018/file/300891a62162b960cf02ce3827bb363c-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2018/file/300891a62162b960cf02ce3827bb363c-Paper.pdf).
- Xiyue Zhang, Benjie Wang, and Marta Kwiatkowska. Provable preimage under-approximation for neural networks. *arXiv preprint arXiv:2305.03686*, 2023.
- Xiangru Zhong. Yolo-benchmark. <https://github.com/xiangrugh/Yolo-Benchmark>, 2023.
- Yuyi Zhong, Quang-Trung Ta, and Siau-Cheng Khoo. Arena: Enhancing abstract refinement for neural network verification. In Cezara Dragoi, Michael Emmi, and Jingbo Wang, editors, *Verification, Model Checking, and Abstract Interpretation*, pages 366–388, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-24950-1.

## A OOD Detection

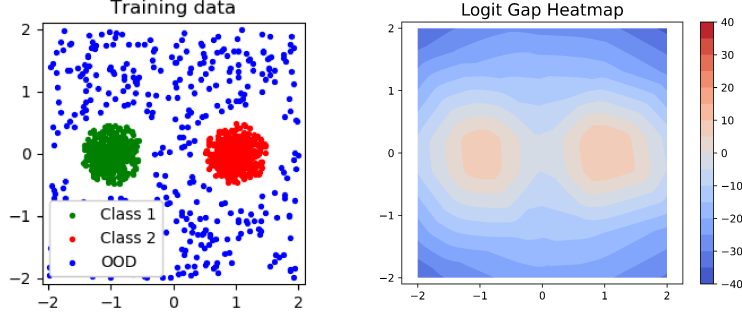


Figure 7: To improve our model’s ability to detect data outside the training distribution (of green and red points), we randomly sample points that are far from every training point. In the right contour, we see that the model’s confidence of in-distribution increases as we approach the centers of the distribution, demonstrating our model is more calibrated.

For the OOD benchmark, we train a neural network on two clusters of data. Standard training would lead to an uncalibrated predictor, we add in OOD data through sampling the input space and rejecting inputs that are too close to any training data point. Training a classifier on this dataset induces the contour on the right, which is a calibrated classifier we can certify properties over. For our experiments, we consider a high-confidence output as 90% confidence of the data being in-distribution.

## B Example

**Example 1.** Consider the toy neural network in Figure 8 under  $\mathcal{X} = [-2, 2]$  and  $\mathcal{S}_{\text{out}} = \{y : 1 \leq y \leq 1.02\}$ . Suppose we wanted to find the tightest  $[l_0^{(1)}, u_0^{(1)}]$  (bounds on  $x_0^{(1)}$ ). If we only enforce the constraints preceding the ReLU (standard bound propagation), we get  $[-2, 2]$ . If we only enforce the constraints following the ReLU, we get  $(-\infty, 1.02]$ . However, when we utilize all of the constraints, we find that the true intermediate bounds are  $[0, 0.01]$ , which is over 100 times tighter than the intersection of the two previous methods (all derivations provided below). Therefore, new techniques are necessary for optimizing intermediate bounds in this setting.

### B.1 Deriving Intermediate Bounds for Example B

**Preceding Constraints.** The linear layer gives us  $x_0^{(1)} = \hat{x}_0^{(0)}$ . Since the only constraint we have is  $\hat{x}_0^{(0)} \in [-2, 2]$ , the tightest intermediate bounds we can derive are  $[-2, 2]$ .

**Following Constraints.** Since the output of a ReLU is non-negative, we know that  $\hat{x}_0^{(1)}, \hat{x}_1^{(1)} \geq 0$ . Since  $x_0^{(2)} = \hat{x}_0^{(1)} + \hat{x}_1^{(1)}$  and the output constraint enforces  $x_0^{(2)} \leq 1.02$ , we derive that  $\hat{x}_0^{(1)} \leq 1.02$ . If we set  $\hat{x}_0^{(1)} = 1.02 - \hat{x}_1^{(1)}$ , we see that  $\hat{x}_0^{(1)}$  can take the entire interval  $[0, 1.02]$ . Since the only constraint we have on  $x_0^{(1)}$  is that  $\text{ReLU}(\hat{x}_0^{(1)}) = x_0^{(1)}$ , our desired interval is the preimage of  $[0, 1.02]$ . This means that  $x_0^{(1)}$  can take any value in the interval  $(-\infty, 1.02]$ .

**All Constraints.** We first note that by the first linear layer, we have  $x_1^{(1)} = x_0^{(1)} + 1$ . Therefore, if  $x_0^{(1)}$  is less than 0, then  $x_1^{(1)}$  is less than 1, which means  $\text{ReLU}(x_0^{(1)}) + \text{ReLU}(x_1^{(1)})$  is less than 1, which contradicts the output constraint. If  $x_0^{(1)}$  is always non-negative, then we have that the output is equivalent to  $\text{ReLU}(x_0^{(1)}) + \text{ReLU}(x_0^{(1)} + 1) = 2x_0^{(1)} + 1$ . Therefore, the output constraint implies  $x_0^{(1)} \leq 0.01$ . Any  $x_0^{(1)} \in [0, 0.01]$  is achievable by setting the input to the desired value.

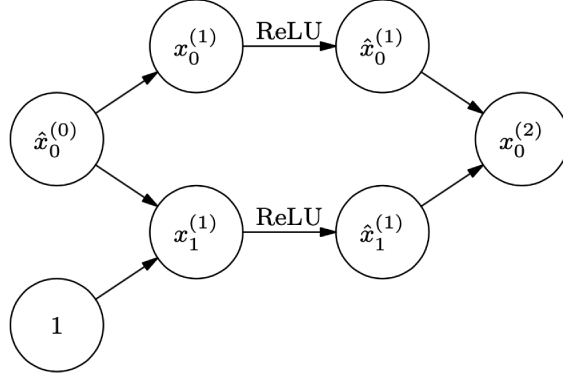


Figure 8: A simple neural network from  $\mathbb{R}$  to  $\mathbb{R}$ . Every node is the sum of its incoming nodes unless the edge is labeled ReLU.

## C Optimizing Intermediate Bounds

We present a generalized theorem which provides a lower bound for any linear combination of  $\hat{\mathbf{x}}^{(0)}$  and  $\mathbf{x}^{(i)}$ . We prove this theorem in Appendix D. We note that by selecting the coefficients of these variables (named  $\mathbf{c}^{(i)}$  for  $i \in \{0, 1, \dots, L-1\}$ ), we recover the two following functionalities which we name  $g_c(\alpha, \gamma)$  and  $g_{i,j}^b(\alpha, \gamma)$  (respectively).

- When the only nonzero coefficients are  $\mathbf{c}^{(0)} = \mathbf{c}$ , we recover Theorem 1. We refer to this bound as  $g_c(\alpha, \gamma)$
- When the only nonzero coefficient is  $c_j^{(i)} = 1$ , we can lower bound  $l_j^{(i)}$  by  $g_{i,j}^{\text{lower}}(\alpha, \gamma)$ . If we set  $c_j^{(i)} = -1$ , we can upper bound  $u_j^{(i)}$  by  $-g_{i,j}^{\text{upper}}(\alpha, \gamma)$ .

With this, we present our generalized theorem.

**Theorem 2** (Lower-bounding combination of neurons). *Given an output set  $\mathcal{S}_{out} = \{\mathbf{y} : \mathbf{H}\mathbf{y} + \mathbf{d} \leq \mathbf{0}\}$  and vector  $\mathbf{c}$ ,  $g(\alpha, \gamma)$  is a lower bound to the linear program in (3) with objective  $\mathbf{c}^{(0)\top} \hat{\mathbf{x}}^{(0)} + \sum_{i=1}^{L-1} \mathbf{c}^{(i)\top} \mathbf{x}^{(i)}$  for  $\mathbf{0} \leq \alpha \leq \mathbf{1}$ ,  $\gamma \geq \mathbf{0}$ , and  $g$  defined via*

$$g(\alpha, \gamma) = \left[ \mathbf{c}^{(0)\top} - \boldsymbol{\nu}^{(1)\top} \mathbf{W}^{(1)} \right]_+ l^{(0)} - \left[ \mathbf{c}^{(0)\top} - \boldsymbol{\nu}^{(1)\top} \mathbf{W}^{(1)} \right]_- \mathbf{u}^{(0)} \\ - \sum_{i=1}^L \boldsymbol{\nu}^{(i)\top} \mathbf{b}^{(i)} + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^{\pm(i)}} \left[ \frac{u_j^{(i)} l_j^{(i)} [\hat{\nu}_j^{(i)}]_+}{u_i^{(j)} - l_i^{(j)}} \right]$$

where every term can be directly recursively computed via

$$\begin{aligned} \mathcal{I}^{-(i)} &= \{j : u_j^{(i)} \leq 0\} \\ \mathcal{I}^{+(i)} &= \{j : l_j^{(i)} \geq 0\} \\ \mathcal{I}^{\pm(i)} &= \{j : l_j^{(i)} < 0 < u_j^{(i)}\} \\ \boldsymbol{\nu}^{(L)} &= -\gamma \\ \hat{\nu}_j^{(i)} &= \boldsymbol{\nu}^{(i+1)\top} \mathbf{W}_{:,j}^{(i+1)} \\ \nu_j^{(i)} &= \begin{cases} \hat{\nu}_j^{(i)} - c_j^{(i)}, & j \in \mathcal{I}^{+(i)} \\ -c_j^{(i)}, & j \in \mathcal{I}^{-(i)} \\ \frac{u_j^{(i)}}{u_i^{(j)} - l_i^{(j)}} [\hat{\nu}_j^{(i)}]_+ - \alpha_j^{(i)} [\hat{\nu}_j^{(i)}]_- - c_j^{(i)}, & j \in \mathcal{I}^{\pm(i)} \end{cases} \end{aligned}$$

## D Dual Derivation

We first incorporate the output constraint  $\mathbf{H}\mathbf{x}^{(L)} + \mathbf{d} \leq \mathbf{0}$  by folding in this linear transformation of the output into the final linear layer of the network, as customary in prior work [Wang et al., 2021, Zhang et al., 2022].

We now prove the generalized theorem as described in C. As a conceptual overview of our proof, we take the Lagrange dual of the linear program to derive an unconstrained optimization problem. From here, we derive constraints that must be satisfied in the max min formulation. These constraints yield the bound propagation procedure we display in Theorem 1.

We start with the convex relaxation.

$$\begin{aligned}
\min_{\mathbf{x}, \hat{\mathbf{x}}} \quad & \mathbf{c}^{(0)\top} \hat{\mathbf{x}}^{(0)} + \sum_{i=1}^{L-1} \mathbf{c}^{(i)\top} \mathbf{x}^{(i)} \\
\text{s.t.} \quad & \mathbf{l}^{(0)} \leq \hat{\mathbf{x}}^{(0)} \leq \mathbf{u}^{(0)} \\
& \mathbf{x}^{(L)} \leq \mathbf{0}; \\
& \mathbf{x}^{(i)} = \mathbf{W}^{(i)} \hat{\mathbf{x}}^{(i-1)} + \mathbf{b}^{(i)}; \quad i \in [L], \\
& \hat{x}_j^{(i)} \geq 0; j \in \mathcal{I}^{\pm(i)} \\
& \hat{x}_j^{(i)} \geq x_j^{(i)}; j \in \mathcal{I}^{\pm(i)} \\
& (u_j^{(i)} - l_j^{(i)}) \hat{x}_j^{(i)} \leq u_j^{(i)} x_j^{(i)} - u_j^{(i)} l_j^{(i)}; j \in \mathcal{I}^{\pm(i)} \\
& \hat{x}_j^{(i)} = x_j^{(i)}; j \in \mathcal{I}^{+(i)} \\
& \hat{x}_j^{(i)} = 0; j \in \mathcal{I}^{-(i)}
\end{aligned}$$

From this, we take the Lagrange dual of most of the constraints. In doing so, we introduce a new dual variable for each constraint. Note that we do not dualize every constraint since they can easily be dealt with later.

$$\begin{aligned}
\min_{\mathbf{x}, \hat{\mathbf{x}}} \max_{\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \boldsymbol{\lambda}} \quad & \mathbf{c}^{(0)\top} \hat{\mathbf{x}}^{(0)} + \sum_{i=1}^{L-1} \mathbf{c}^{(i)\top} \mathbf{x}^{(i)} + \boldsymbol{\gamma}^\top \mathbf{x}^{(L)} + \sum_{i=1}^L \boldsymbol{\nu}^{(i)\top} \left( \mathbf{x}^{(i)} - \mathbf{W}^{(i)} \hat{\mathbf{x}}^{(i-1)} - \mathbf{b}^{(i)} \right) \\
& + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^{\pm(i)}} \left[ \mu_j^{(i)} \left( -\hat{x}_j^{(i)} \right) + \tau_j^{(i)} \left( x_j^{(i)} - \hat{x}_j^{(i)} \right) + \lambda_j^{(i)} \left( (u_j^{(i)} - l_j^{(i)}) \hat{x}_j^{(i)} - u_j^{(i)} x_j^{(i)} + u_j^{(i)} l_j^{(i)} \right) \right] \\
\text{s.t.} \quad & \mathbf{l}^{(0)} \leq \hat{\mathbf{x}}^{(0)} \leq \mathbf{u}^{(0)}; \quad \hat{x}_j^{(i)} = 0, j \in \mathcal{I}^{-(i)}; \quad \hat{x}_j^{(i)} = x_j^{(i)}, j \in \mathcal{I}^{+(i)} \\
& \boldsymbol{\mu} \geq \mathbf{0}; \quad \boldsymbol{\tau} \geq \mathbf{0}; \quad \boldsymbol{\gamma} \geq \mathbf{0}; \quad \boldsymbol{\lambda} \geq \mathbf{0}
\end{aligned}$$

Since we took the dual of a linear program, the solution to the min-max optimization is equivalent to the solution of the max-min optimization by strong duality. Therefore, we can solve the following program (equivalent up to rearrangement).

$$\begin{aligned}
& \max_{\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \boldsymbol{\lambda}} \min_{\mathbf{x}, \hat{\mathbf{x}}} \left( \boldsymbol{\nu}^{(L)} + \boldsymbol{\gamma} \right)^\top \mathbf{x}^{(L)} + \left( \mathbf{c}^{(0)\top} - \boldsymbol{\nu}^{(1)\top} \mathbf{W}^{(1)} \right) \hat{\mathbf{x}}^{(0)} \\
& + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^{+(i)}} \left( \nu_j^{(i)} - \boldsymbol{\nu}^{(i+1)\top} \mathbf{W}_{:,j}^{(i+1)} + c_j^{(i)} \right) x_j^{(i)} + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^{-(i)}} \left( \nu_j^{(i)} + c_j^{(i)} \right) x_j^{(i)} \\
& + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^{\pm(i)}} \left[ \left( \nu_j^{(i)} + \tau_j^{(i)} - \lambda_j^{(i)} u_j^{(i)} + c_j^{(i)} \right) x_j^{(i)} \right. \\
& \left. + \left( -\boldsymbol{\nu}^{(i+1)\top} \mathbf{W}_{:,j}^{(i+1)} - \mu_j^{(i)} - \tau_j^{(i)} + (u_j^{(i)} - l_j^{(i)}) \lambda_j^{(i)} \right) \hat{x}_j^{(i)} \right] \\
& - \sum_{i=1}^L \boldsymbol{\nu}^{(i)\top} \mathbf{b}^{(i)} + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^{\pm(i)}} \lambda_j^{(i)} u_j^{(i)} l_j^{(i)} \\
& \text{s.t. } \mathbf{l}^{(0)} \leq \hat{\mathbf{x}}^{(0)} \leq \mathbf{u}^{(0)} \\
& \quad \boldsymbol{\mu} \geq 0; \quad \boldsymbol{\tau} \geq 0; \quad \boldsymbol{\gamma} \geq 0; \quad \boldsymbol{\lambda} \geq 0
\end{aligned}$$

To minimize  $(\mathbf{c}^{(0)\top} - \boldsymbol{\nu}^{(1)\top} \mathbf{W}^{(1)}) \hat{\mathbf{x}}^{(0)}$  subject to  $\mathbf{l}^{(0)} \leq \hat{\mathbf{x}}^{(0)} \leq \mathbf{u}^{(0)}$ , we can consider the choice we must make in each dimension. If the  $j$ th entry of the coefficient is positive, we should set  $\hat{x}_j^{(0)} = l_j^{(0)}$ . Otherwise, we should set  $\hat{x}_j^{(0)} = u_j^{(0)}$ .

$$\begin{aligned}
& \max_{\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \boldsymbol{\lambda}} \min_{\mathbf{x}, \hat{\mathbf{x}}} \left( \boldsymbol{\nu}^{(L)} + \boldsymbol{\gamma} \right)^\top \mathbf{x}^{(L)} + \left[ \mathbf{c}^{(0)\top} - \boldsymbol{\nu}^{(1)\top} \mathbf{W}^{(1)} \right]_+ \mathbf{l}^{(0)} - \left[ \mathbf{c}^{(0)\top} - \boldsymbol{\nu}^{(1)\top} \mathbf{W}^{(1)} \right]_- \mathbf{u}^{(0)} \\
& + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^{+(i)}} \left( \nu_j^{(i)} - \boldsymbol{\nu}^{(i+1)\top} \mathbf{W}_{:,j}^{(i+1)} + c_j^{(i)} \right) x_j^{(i)} + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^{-(i)}} \left( \nu_j^{(i)} + c_j^{(i)} \right) x_j^{(i)} \\
& + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^{\pm(i)}} \left[ \left( \nu_j^{(i)} + \tau_j^{(i)} - \lambda_j^{(i)} u_j^{(i)} + c_j^{(i)} \right) x_j^{(i)} \right. \\
& \left. + \left( -\boldsymbol{\nu}^{(i+1)\top} \mathbf{W}_{:,j}^{(i+1)} - \mu_j^{(i)} - \tau_j^{(i)} + (u_j^{(i)} - l_j^{(i)}) \lambda_j^{(i)} \right) \hat{x}_j^{(i)} \right] \\
& - \sum_{i=1}^L \boldsymbol{\nu}^{(i)\top} \mathbf{b}^{(i)} + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^{\pm(i)}} \lambda_j^{(i)} u_j^{(i)} l_j^{(i)} \\
& \text{s.t. } \boldsymbol{\mu} \geq 0; \quad \boldsymbol{\tau} \geq 0; \quad \boldsymbol{\gamma} \geq 0; \quad \boldsymbol{\lambda} \geq 0
\end{aligned}$$

From here, we note that the variables  $\mathbf{x}$  or  $\hat{\mathbf{x}}$  are unconstrained variables. Therefore, if any of their coefficients are nonzero, the inner minimization can immediately drive its value to  $-\infty$ . As such, the outer maximization must set all of these coefficients to zero. Therefore, we can derive constraints from this restructured optimization and remove the free variables  $\mathbf{x}, \hat{\mathbf{x}}$ .

$$\begin{aligned}
& \max_{\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \boldsymbol{\lambda}} \quad \left[ \mathbf{c}^{(0)\top} - \boldsymbol{\nu}^{(1)\top} \mathbf{W}^{(1)} \right]_+ \mathbf{l}^{(0)} - \left[ \mathbf{c}^{(0)\top} - \boldsymbol{\nu}^{(1)\top} \mathbf{W}^{(1)} \right]_- \mathbf{u}^{(0)} - \sum_{i=1}^L \boldsymbol{\nu}^{(i)\top} \mathbf{b}^{(i)} \\
& \quad + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^{\pm(i)}} \lambda_j^{(i)} u_j^{(i)} l_j^{(i)} \\
& \text{s.t.} \quad \boldsymbol{\nu}^{(L)} = -\boldsymbol{\gamma} \\
& \quad \nu_j^{(i)} = \boldsymbol{\nu}^{(i+1)\top} \mathbf{W}_{:,j}^{(i+1)} - c_j^{(i)}, j \in \mathcal{I}^{+(i)} \\
& \quad \nu_j^{(i)} = -c_j^{(i)}, j \in \mathcal{I}^{-(i)} \\
& \quad \nu_j^{(i)} = \lambda_j^{(i)} u_j^{(i)} - \tau_j^{(i)} - c_j^{(i)}, j \in \mathcal{I}^{\pm(i)} \\
& \quad \boldsymbol{\nu}^{(i+1)\top} \mathbf{W}_{:,j}^{(i+1)} = (u_j^{(i)} - l_j^{(i)}) \lambda_j^{(i)} - \left( \mu_j^{(i)} + \tau_j^{(i)} \right), j \in \mathcal{I}^{\pm(i)} \\
& \quad \boldsymbol{\mu} \geq 0; \quad \boldsymbol{\tau} \geq 0; \quad \boldsymbol{\gamma} \geq 0; \quad \boldsymbol{\lambda} \geq 0
\end{aligned}$$

For the following, we define  $\hat{\nu}_j^{(i)} = \boldsymbol{\nu}^{(i+1)\top} \mathbf{W}_{:,j}^{(i+1)}$ . We note that since the upper and lower bounds of the neuron relaxation can not be tight simultaneously, at least one of  $(u_j^{(i)} - l_j^{(i)}) \lambda_j^{(i)}$  and  $\mu_j^{(i)} + \tau_j^{(i)}$  must be non-zero. Therefore, we can write them as  $(u_j^{(i)} - l_j^{(i)}) \lambda_j^{(i)} = [\hat{\nu}_j^{(i)}]_+$  and  $\mu_j^{(i)} + \tau_j^{(i)} = [\hat{\nu}_j^{(i)}]_-$ . We can then use the fact that  $\tau_j^{(i)}$  lies in the interval 0 and  $\hat{\nu}_j^{(i)}$  to get the following bound propagation procedure.

$$\begin{aligned}
& \max_{\boldsymbol{\nu}, \boldsymbol{\alpha}, \boldsymbol{\gamma}} \quad \left[ \mathbf{c}^{(0)\top} - \boldsymbol{\nu}^{(1)\top} \mathbf{W}^{(1)} \right]_+ \mathbf{l}^{(0)} - \left[ \mathbf{c}^{(0)\top} - \boldsymbol{\nu}^{(1)\top} \mathbf{W}^{(1)} \right]_- \mathbf{u}^{(0)} - \sum_{i=1}^L \boldsymbol{\nu}^{(i)\top} \mathbf{b}^{(i)} \\
& \quad + \sum_{i=1}^{L-1} \sum_{j \in \mathcal{I}^{\pm(i)}} \left[ \frac{u_j^{(i)} l_j^{(i)} [\hat{\nu}_j^{(i)}]_+}{u_i^{(j)} - l_i^{(j)}} \right] \\
& \text{s.t.} \quad \boldsymbol{\nu}^{(L)} = -\boldsymbol{\gamma} \\
& \quad \nu_j^{(i)} = \boldsymbol{\nu}^{(i+1)\top} \mathbf{W}_{:,j}^{(i+1)} - c_j^{(i)}, j \in \mathcal{I}^{+(i)} \\
& \quad \nu_j^{(i)} = -c_j^{(i)}, j \in \mathcal{I}^{-(i)} \\
& \quad \hat{\nu}_j^{(i)} = \boldsymbol{\nu}^{(i+1)\top} \mathbf{W}_{:,j}^{(i+1)} \\
& \quad \nu_j^{(i)} = \frac{u_j^{(i)}}{u_i^{(j)} - l_i^{(j)}} [\hat{\nu}_j^{(i)}]_+ - \alpha_j^{(i)} [\hat{\nu}_j^{(i)}]_- - c_j^{(i)}, j \in \mathcal{I}^{\pm(i)} \\
& \quad \alpha_j^{(i)} \in [0, 1]; \quad \boldsymbol{\gamma} \geq 0
\end{aligned}$$

In this program,  $\alpha_j^{(i)}$  are optimizable parameters controlling the relaxation of neuron  $j$  in layer  $i$ , similar to the ones appearing in Xu et al. 2021.  $\boldsymbol{\gamma}$ , as discussed in the body of the paper, is the parameter which enforces the output constraint throughout this entire bound propagation procedure.

## E Connection to Forward Verification for Intermediate Bounds

We consider the general objective presented in C and aim to minimize  $\mathbf{c}^{(0)\top} \hat{\mathbf{x}}^{(0)} + \sum_{i=1}^{L-1} \mathbf{c}^{(i)\top} \mathbf{x}^{(i)}$

$$\begin{aligned}
& \min_{\mathbf{x}} \quad \mathbf{c}^{(0)\top} \hat{\mathbf{x}}^{(0)} + \sum_{i=1}^{L-1} \mathbf{c}^{(i)\top} \mathbf{x}^{(i)} \\
& \text{s.t.} \quad \mathbf{x} \in \mathcal{X}; \quad \mathbf{H}\mathbf{f}(\mathbf{x}) + \mathbf{d} \leq 0
\end{aligned}$$

If we take the dual of the constraint, then we get the program

$$\begin{aligned}
\min_{\mathbf{x}} \quad & \mathbf{c}^{(0)\top} \hat{\mathbf{x}}^{(0)} + \sum_{i=1}^{L-1} \mathbf{c}^{(i)\top} \mathbf{x}^{(i)} + \gamma (\mathbf{H}f(\mathbf{x}) + \mathbf{d}) \\
\text{s.t.} \quad & \mathbf{x} \in \mathcal{X}; \quad \gamma \geq 0
\end{aligned}$$

We note that the objective here can be expressed as a neural network with residual stream  $f(\mathbf{x})$  and a skip connection with linear weight  $\mathbf{c}^{(i)}$  from the pre-activations of layer  $i$  to the final output for every layer. In theory, we could construct this neural network and directly pass it to a forward verification tool which could iteratively tighten all bounds for a solution.

## F Implementation

As stated in Algorithm 1, INVPROP first initializes bounds for all layers using some computationally cheap technique. Based on the input bounds given by  $\mathcal{X}$ , we first compute intermediate bounds using interval propagation and then tighten them based on the reverse symbolic interval propagation (RSIP) technique Singh et al. [2019]. While INVPROP will iteratively tighten those bounds over time, we found RSIP to reduce the total necessary runtime significantly compared to an initialization based solely on interval propagation. We initialize all  $\alpha$  with 0.5 and all  $\gamma$  with 0.025.

The optimization is performed for all lower and upper bounds of all neurons in each layer in parallel, starting with the last layer and moving forward to the input layer. The bounds of the cutting hyperplanes are optimized last, together with the bounds on the input neurons. The improvements on  $\mathbf{c}^\top \mathbf{x}$  are measured every 10th iteration. Before doing so, the bounds of the hyperplanes are tightened for 10 extra steps to improve their precision. We detect convergence by monitoring  $\mathbf{c}^\top \mathbf{x}$ . If successive iterations see minimal improvement, we stop or branch on the input space. All cutting hyperplanes are evenly distributed to maximize their information gain. For 40 hyperplanes in a 2D input space, we rotate each plane by  $9^\circ$ . All reported runtimes under 10 minutes are computed as the average of five runs.

### F.1 Encoding Non-Linear Constraints

To support non-linear output sets, such as the maximum operation used in the OOD example in Section 4.3, the non-linearity needs to be encoded, such that it can be expressed as linear constraints over the modified network. We rewrite  $\max(y_1, y_2) = \max(y_1 - y_2, 0) + y_2$  and add an additional ReLU layer for this operation. Note that to pass  $y_2$  through this layer without modifying it, one can either write  $y_2 = \max(y_2, 0) - \max(-y_2, 0)$ , or  $y_2 = \max(y_2 - M, 0) + M$  where  $M$  is the lower bound of all possible  $y_2$ . We find that avoiding the additional ReLU relaxations that would occur for the first approach is beneficial for the optimization and compute a lower bound of  $M$  using interval propagation. As  $y_3$  (for the OOD class) should not be changed by this operation either, we apply the same trick of subtracting and adding its lower bound.

### F.2 Control Benchmark Encoding

We encode the entire control formula  $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$  as one feedforward network by encoding the residual connection as regular fully connected layers. To this end, we use the same technique described in Section F.1 to shift the bounds into the positive regime, feed them forward and then shift them back.

To compute  $\mathcal{S}_{\text{over}}$  for a timestep  $t > 1$ , we first stack the network  $t$  times, then simplify it by merging consecutive linear layers. All bounds of layers not affected by this merging that also appear in the network for  $t - 1$  timesteps are reused. Their bounds are already tight enough and are not optimized further. Note that this is different than using the previous  $\mathcal{S}_{\text{over}}$  as the new target region and using an unstacked network: All bounds of the new layers are still optimized w.r.t. the precise target  $\mathcal{S}_{\text{out}}$ . Therefore, we do not suffer from accumulating inaccuracies.

## G Hardware

For the control benchmark, all experiments were performed on a Dual Xeon Gold 6138.

For the OOD benchmark, all experiments were performed on an Intel Xeon Platinum 8160 processor using 8 cores and 40GB RAM, as well as a V100-SXM2 GPU.

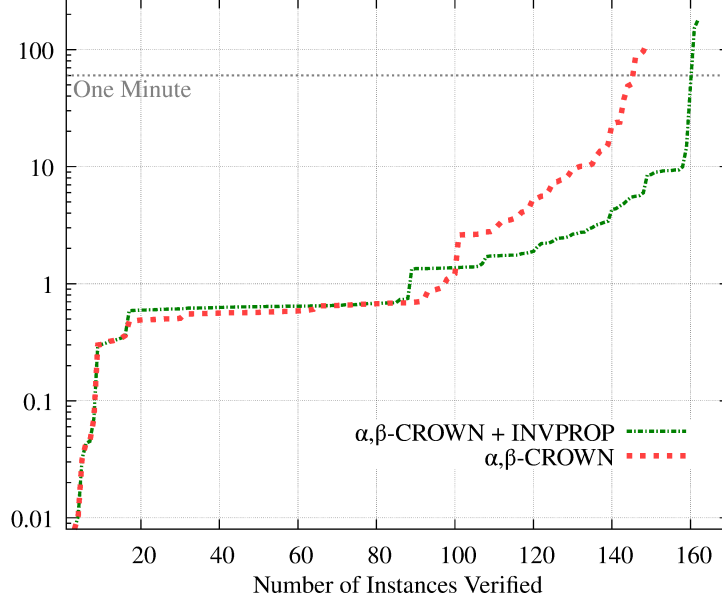


Figure 9: Detailed runtime comparison of  $\alpha, \beta$ -CROWN and  $\alpha, \beta$ -CROWN + INVPROP. Except for a few instances,  $\alpha, \beta$ -CROWN extended with INVPROP can verify the properties faster, and can prove robustness for some instances that cause a timeout for pure  $\alpha, \beta$ -CROWN. The timeout per instance is 5 minutes.

For the MNIST and YOLO robustness verification benchmarks, an AWS instance of type g5.2xlarge was used, which is equipped with a AMD EPYC 7R32 processor with 8 cores, 32GB RAM, as well as an A10G GPU.

## H Robustness Verification

For the comparison in Section 4.2, we have implemented the concept of INVPROP in  $\alpha, \beta$ -CROWN. Figure 9 compares the performance of  $\alpha, \beta$ -CROWN and  $\alpha, \beta$ -CROWN + INVPROP on the MNIST benchmark defined by Yang et al. [2021b].