Distillation based Robustness Verification with PAC Guarantees

Patrick Indri [*] [1] **Peter Blohm** [*] [1] **Anagha Athavale** [1] **Ezio Bartocci** [1] **Georg Weissenbacher** [1] **Matteo Maffei** [1] **Dejan Nickovic** [2] **Thomas Gärtner** [1] **Sagar Malhotra** [1]

Abstract

We present a distillation based approach to verify the robustness of any Neural Network (NN). Conventional formal verification methods cannot tractably assess the global robustness of real-world NNs. To address this, we take advantage of a gradient-aligned distillation framework to transfer the robustness properties from a larger teacher network to a smaller student network. Given that the student NN can be formally verified for global robustness, we theoretically investigate how this guarantee can be transferred to the teacher NN. We draw from ideas in learning theory to derive a sample complexity for the distillation procedure that enables PAC-guarantees on the global robustness of the teacher network.

1. Introduction

We use knowledge distillation to provide robustness guarantees for a larger teacher Neural Network (NN) by formally verifying the robustness of a smaller student distilled from it. NN verification is an important task, especially when NNs are used in safety-critical applications like autonomous driving or medical diagnosis. However, small perturbations to the input may lead NNs to significantly change their prediction and thus to misclassify (Goodfellow et al., 2015; Kurakin et al., 2018). A potential solution to these challenges is to use robust NNs, that is, NNs that do not change their prediction under small perturbations to the input. Various approaches to assess the robustness of NNs have been proposed in the literature (Wu et al., 2020; Webb et al., 2019). However, methods that provide formal robustness guarantees are limited to small scale NNs with few parameters (Seshia et al., 2018; Katz et al., 2017; Huang et al., 2017; Gopinath et al., 2018).

In this paper, we investigate robustness certification of large-scale NNs using existing formal verification tools. To this end, we investigate a gradient-aligned knowledge distillation procedure to distill a larger teacher NN into a smaller NN that can be tractably verified. Knowledge distillation with gradient alignment (Chan et al., 2020; Shao et al., 2021; Lee et al., 2023) promotes a transfer of local robustness properties from the teacher to the student NN. Shao et al. (2021) show that if teacher and student agree on the predictions and the gradients on the data points used for distillation, then they have similar local robustness properties at these points. Gradient-aligned distillation by itself does not, however, necessarily preserve global robustness. We nevertheless formally demonstrate that if the distillation took place on a large enough sample and the student is verified to be globally robust, then the teacher is globally robust with a high probability. We also investigate situations where the distillation procedure is not perfect. To give robustness guarantees in such scenarios, we derive new conditions based on the empirically observed discrepancy of the gradients. We provide initial empirical evidence for our distillation based robustness verification procedure. Our results show that distillation, if performed successfully, effectively preserves robustness properties between teacher and the student NNs. Furthermore, when the distillation procedure is not successful, we show that clear empirical markers can identify this. Hence, the robustness of the student NN can be used as a conservative guarantee for the robustness of the teacher NN.

Problem statement Let $f_T(\mathbf{x})$ be a classifier for which we want to provide global robustness guarantees. We consider a setting where $f_T(\mathbf{x})$ is too large to be verified using formal methods. We distill $f_T(\mathbf{x})$ into a smaller classifier $f_S(\mathbf{x})$ while preserving robustness properties. $f_S(\mathbf{x})$ is small enough to be formally verified and we want to use the information obtained from the verification of $f_S(\mathbf{x})$ to provide robustness guarantees on the larger $f_T(\mathbf{x})$.

2. Related work

The use of NN models in safety-critical applications such as in medical diagnosis (Amato et al., 2013) or in self-driving cars (Rao & Frtunikj, 2018) raises concerns about their safety. Research efforts have thus been made towards the

[*] Equal contribution [1] TU Wien, Austria [2] ISTA, Austria. Correspondence to: Patrick Indri <patrick.indri@tuwien.ac.at>, Peter Blohm <peter.blohm@tuwien.ac.at>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

automated verification of NNs, that is, towards providing formal guarantees of their behavior (Chen et al., 2021) using formal verification techniques (Katz et al., 2019). Verifying the robustness of a NN can guarantee that a network provides reliable prediction even when data is perturbed (Casadio et al., 2022; Meng et al., 2022). Conventional formal verification methods require significant computational resources. Hence, recent approaches have proposed techniques to give *approximate* (Wu et al., 2020) or *statistical* (Webb et al., 2019) robustness guarantees.

The machine learning community has made significant efforts to devise techniques to train NNs which are certifiably robust (Li et al., 2023). Specifically, knowledge distillation has been proposed as a technique to transfer robustness between models (Chan et al., 2020; Shao et al., 2021; Lee et al., 2023), to boost robustness (Vaishnavi et al., 2022; Huang et al., 2023), as well as to provide effective defenses against adversarial attacks (Papernot et al., 2016).

3. Preliminaries

Given a vector $\mathbf{x} \in \mathbb{R}^m$, we use $\|\mathbf{x}\|$ to denote its L_2 norm. Let $B_r(\mathbf{x}) := \{\mathbf{x}' \mid \|\mathbf{x}' - \mathbf{x}\| \leq r\}$ denote the closed L_2-ball centered at $\mathbf{x}$ with radius $r > 0$. Given a function f, we use $\nabla_{\mathbf{x}} f$ to denote its gradient with respect to $\mathbf{x}$. We use $\sigma(x)$ to denote the sigmoid function $\frac{1}{1+e^x}$.

In a classification task, we consider a (training) dataset D consisting of point-label pairs $(\mathbf{x}, y)$. For a task with $n \in \mathbb{N}$ classes, each point $\mathbf{x} \in \mathbb{R}^m$ is associated with a label $y \in [1, n]$ With a slight abuse of notation, we write $\mathbf{x} \in D_{\mathbf{x}}$ or $(\mathbf{x}, y) \in D$ to refer respectively to the points or the point-label pairs of D.

Robustness A classifier f is *robust* if its prediction $f(\mathbf{x})$ does not change under small perturbations around $\mathbf{x}$. Robustness thus preserves the performance of a classifier in the event of small input perturbations. A local notion of robustness can be defined as follows.

Definition 3.1 (Local δ-robustness, Leino et al. (2021); Athavale et al. (2024))**.** A classifier $f(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}^n$ is locally *δ-robust* in $\mathbf{x}$ if

$$\forall\, \mathbf{x}' \in B_\delta(\mathbf{x}) \quad \arg\max f(\mathbf{x}') = \arg\max f(\mathbf{x}), \tag{1}$$

where, in a classification setting, the $\arg\max$ returns the index corresponding to the predicted class.

A natural definition of *global* δ-robustness would require local δ-robustness for all possible inputs. However, requiring robustness on all possible inputs, in spaces like $\mathbb{R}^m$, would only permit classifiers with constant prediction to be globally robust. Hence, a more useful notion of global robustness is to require a classifier to be robust on all points in certain regions of interest. For instance, for all inputs $\mathbf{x} \in D_{\mathbf{x}}$, we will want the classifier to be robust only for the regions of high *confidence* (Leino et al., 2021; Athavale et al., 2024). Similarly to Athavale et al. (2024) we assume that the predictions of our classifier (usually a NN) are the result of a softmax function. We use this to define *confidence* as follows:

Definition 3.2 (Confidence)**.** A classifier $f(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}^n$ is κ-confident for some $\mathbf{x} \in \mathbb{R}^m$ if

$$\max(\text{softmax}(f(\mathbf{x})) \geq \kappa, \tag{2}$$

where the $\max$ is computed with respect to the n classes.

Definition 3.3 (Global δ-robustness)**.** A classifier $f(\mathbf{x})$ is *globally δ-robust* in a set $D_{\mathbf{x}}$ with respect to a confidence threshold κ if it is locally δ-robust for all $\mathbf{x} \in D_{\mathbf{x}}$ for which it is κ-confident.

Distillation First proposed by Buciluă et al. (2006), knowledge distillation (Wang & Yoon, 2022; Hinton et al., 2015) entails the training of a student network under the supervision of a teacher network. The student is trained by minimizing the difference in the logits produced by the teacher and the student model. Let $f_T(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}^n$ be the teacher model and $f_S(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}^n$ be the student model. A typical knowledge distillation setting may consider a distillation loss $\mathcal{L}_{\text{KD}}$ proportional to sum of (i) the cross-entropy loss of the student $\mathcal{L}_{\text{CE}}$ and (ii) the KL-divergence loss between the student and the teacher models $\mathcal{L}_{\text{KL}}$. Hence, we have that

$$\mathcal{L}_{\text{KD}}(\mathbf{x}, y) \propto \mathcal{L}_{\text{CE}}(f_S(\mathbf{x}), y) + \mathcal{L}_{\text{KL}}(f_S(\mathbf{x}), f_T(\mathbf{x})). \tag{3}$$

We refer the reader to, e.g., Wang & Yoon (2022) for a survey on knowledge distillation.

NN Verification Automated verification techniques can be used to provide formal guarantees about the robustness of a NN. We assume that we can use a formal verification tool, such as the one in Athavale et al. (2024), to check the global robustness of a small NN. Our goal is to investigate how robustness guarantees from such a tool on a small student network can be used to infer robustness guarantees on the larger teacher NN.

ϵ-nets Given a probability distribution, an ϵ-net is a representative set of points such that all high probability density regions of the distribution are intersected by the set (Haussler & Welzl, 1986; Mustafa & Varadarajan, 2017).

Definition 3.4 (Range space)**.** Let $\mathcal{X}$ be a (possibly infinite) set of *points* and $\mathcal{R}$ a family of subsets of $\mathcal{X}$ called *ranges*. A range space is defined as the pair $(\mathcal{X}, \mathcal{R})$.

Definition 3.5 (Shattering)**.** We say a subset $S \subseteq \mathcal{X}$ is *shattered* by $\mathcal{R}$ if for any $S' \subseteq S$ there exists an $R \in \mathcal{R}$ such that $S' = R \cap S$.

Definition 3.6 (Vapnik–Chervonenkis (VC) dimension)**.** The VC dimension of a range space $(\mathcal{X}, \mathcal{R})$ is the maximum cardinality of a set $S \subseteq \mathcal{R}$ that is shattered by $\mathcal{R}$.

Definition 3.7 (ϵ-net, Haussler & Welzl (1986))**.** Let $(\mathcal{X}, \mathcal{R})$ be a range space and $\mathcal{D}_{\mathcal{X}}$ be a probability distribution on $\mathcal{X}$. A set $N \subseteq \mathcal{X}$ is an ϵ-net for $\mathcal{X}$ with respect to $\mathcal{D}_{\mathcal{X}}$ if for every set $R \in \mathcal{R}$ such that $\Pr_{\mathcal{D}_{\mathcal{X}}}(R) \geq \epsilon$, the set R contains at least one point from N, i.e.,

$$\forall R \in \mathcal{R},\ Pr_{\mathcal{D}_{\mathcal{X}}}(R) \geq \epsilon \implies R \cap N \neq \emptyset. \tag{4}$$

A set of points $S \in \mathcal{X}$ is thus an ϵ-net for a space $\mathcal{X}$ if it intersects any range $R \in \mathcal{R}$ of sufficient probability.

Obtaining an ϵ-net of optimal size is intractable in general. However, the following theorem shows that a sufficiently large number of i.i.d. samples form an ϵ-net with high probability.

Theorem 3.8 (ϵ-nets from i.i.d. samples, Mitzenmacher & Upfal (2017))**.** *Let $(\mathcal{X}, \mathcal{R})$ be a range space with VC dimension d and let $\mathcal{D}$ be a probability distribution on $\mathcal{X}$. For any $0 < \eta, \epsilon \leq \frac{1}{2}$, an i.i.d. sample from $\mathcal{D}$ of size s is an ϵ-net for $\mathcal{X}$ with probability at least $1 - \eta$ if*

$$s \geq \frac{8d}{\epsilon} \ln \frac{16d}{\epsilon} + \frac{4}{\epsilon} \ln \frac{2}{\eta}. \tag{5}$$

4. Robustness guarantees

In this section, we describe our approach to provide global robustness guarantees for a large NN model which cannot be tractably verified using formal verification tools. We first revisit the local robustness preserving distillation approach as presented by Shao et al. (2021). As the distillation procedure may not work perfectly in real-world applications, we derive relaxed local robustness guarantees that can be obtained from an imperfect distillation procedure. We then provide a statistical argument to provide PAC-guarantees on global robustness of the teacher model depending on the sample size we use for distillation.

4.1. Transferring robustness properties via distillation

Chan et al. (2020) and Shao et al. (2021) showed that if two models have the same predictions and gradients on a set of data points, then they share similar local robustness properties around these points. From a given teacher model, we obtain a student model with similar local behavior through knowledge distillation and then infer the local robustness properties of the teacher model from those of its student.

Definition 4.1 (Gradient-aligned distillation, adapted from Shao et al. (2021))**.** Gradient-aligned distillation is a training procedure that takes a fixed teacher model f_T and trains a student model f_S to minimize the distillation loss $\mathcal{L}_{\text{G}}$. Let $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{KL}}$ denote the cross-entropy loss and the KL-divergence loss respectively. Let $\lambda_{\text{CE}}, \lambda_{\text{KL}}, \lambda_{\text{G}}$, and τ be some hyperparameters. The gradient-aligned distillation loss $\mathcal{L}_G$ is then defined as:

$$\begin{aligned}\mathcal{L}_{\text{G}}(\mathbf{x}, y) = \ &\lambda_{\text{CE}}\mathcal{L}_{\text{CE}}(f_S(\mathbf{x}), y) \ + \\ &\lambda_{\text{KL}}\tau^2\mathcal{L}_{\text{KL}}\left(\frac{f_S(\mathbf{x})}{\tau}, \frac{f_T(\mathbf{x})}{\tau}\right) \ + \\ &\lambda_{\text{G}}\|\nabla_{\mathbf{x}}\mathcal{L}_{\text{CE}}(f_S(\mathbf{x}), y) - \nabla_{\mathbf{x}}\mathcal{L}_{\text{CE}}(f_T(\mathbf{x}), y)\|.\end{aligned} \tag{6}$$

Similarly to Shao et al. (2021), we assume that we can obtain a *perfect* student for which

$$\forall\, \mathbf{x} \in D_{\mathbf{x}},\ \mathcal{L}_{\text{G}}(\mathbf{x}, \arg\max(f_T(\mathbf{x}))) = 0. \tag{7}$$

Where $D_{\mathbf{x}}$ denotes the unlabeled dataset used for distillation and the labels are produced by teacher. Additionally, we assume that: (i) both f_T and f_S are *locally linear* (Lee et al., 2019; Sattelberg et al., 2023), and (ii) the areas of certified robustness (alternatively, that the output of any training input) falls into one of these locally linear areas. The local linearity assumption (i) is justified for networks with piece-wise linear or, in general, piece-wise affine activation functions (Croce et al., 2019). As any such network represents a piece-wise linear/affine function and vice versa (Arora et al., 2018, Theorem 2.1). Even though for a typical network the linear regions may be small, assumption (ii) is justified if proper regularization is used, as empirically shown by Croce et al. (2019) who obtain large linear regions for ReLU networks. The main focus of this work is to verify the robustness of a NN rather than to train a robust one. Hence, we assume that proper precautions have been met during the training of the teacher to satisfy (ii).

Proposition 4.2 (Robustness of the teacher model)**.** *Consider a teacher model f_T and a perfect student model f_S obtained through the distillation process described in Definition 4.1. $\forall \mathbf{x} \in D_{\mathbf{x}}$, if $f_S(\mathbf{x})$ is locally δ-robust in $\mathbf{x}$, then $f_T(\mathbf{x})$ is locally δ-robust in $\mathbf{x}$.*

Perfect gradient alignment is necessary to transfer robustness properties from the student to the teacher. However, we prove a weaker guarantee for imperfect alignments.

Proposition 4.3 (Imperfect gradient alignment)**.** *Consider a teacher model $f_T(\mathbf{x})$ and a student model $f_S(\mathbf{x})$ such that $\forall\ (\mathbf{x}, y) \in D\ f_S(\mathbf{x}) = f_T(\mathbf{x}) = y$. If $f_S(\mathbf{x})$ is locally δ-robust in $\mathbf{x}$, then $f_T(\mathbf{x})$ is locally δ_T-robust in $\mathbf{x}$, where*

$$\delta_T = \|\nabla_{\mathbf{x}} f_S(\mathbf{x})\| \|\nabla_{\mathbf{x}} f_T(\mathbf{x})\|^{-1}\delta. \tag{8}$$

4.2. PAC global robustness

We provide a *Probably Approximately Correct* (PAC) guarantee (Valiant, 1984; Haussler & Warmuth, 2018) on the global robustness of a classifier. We use the notion of ϵ-net described in Section 3 to obtain global robustness guarantees with high probability from local robustness.

Consider the Euclidean space $\mathcal{X} = \mathbb{R}^m$ and a set of ranges containing all metric balls of some fixed radius δ: $\mathcal{R} = \{B_\delta(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^m\}$. The VC-dimension of this range space corresponds to the VC-dimension of the set of all closed balls of arbitrary radius in $\mathbb{R}^m$ and is equal to $m+1$ (Dudley, 1979). Let $\mathcal{D}_\mathcal{X}$ be a continuous distribution over $\mathcal{X}$ and let $\epsilon \in [0, 1]$, $\delta \in \mathbb{R}_{>0}$. Let $\mathcal{B}_{\delta,\epsilon}$ denote the set of closed balls of radius δ with a probability mass of at least ϵ, i.e.,

$$\mathcal{B}_{\delta,\epsilon} = \{B_\delta(\mathbf{x}) \in \mathcal{R} \mid \Pr_{\mathcal{D}_\mathcal{X}}(B_\delta(\mathbf{x})) \geq \epsilon\}. \tag{9}$$

Theorem 4.4. *Let ϵ, $\eta \in [0, 1]$, $\delta \in \mathbb{R}_{>0}$. If a classifier $f(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}^n$ is locally 3δ-robust in*

$$s \geq \frac{8(m+1)}{\epsilon} \ln \frac{16(m+1)}{\epsilon} + \frac{4}{\epsilon} \ln \frac{2}{\eta} \tag{10}$$

points sampled from $\mathcal{D}_\mathcal{X}$, then it is globally δ-robust in any point covered by a ball in $\mathcal{B}_{\delta,\epsilon}$ with probability at least $1-\eta$.

Refer to Appendix A for the proof. Theorem 4.4 thus suggests that, given enough samples on which a classifier f is locally robust, we can guarantee with high probability that the classifier is globally robust in any point covered by a ball in $\mathcal{B}_{\delta,\epsilon}$.

Note that we make no assumption on the distribution $\mathcal{D}_\mathcal{X}$. In fact, the guarantee obtained in Theorem 4.4 can be adapted to any distribution which is *close* to $\mathcal{D}_\mathcal{X}$. Consider the total variation distance between two probabilities as $\mathrm{TV}(\mathcal{D}_\mathcal{X}, \mathcal{D}'_\mathcal{X}) = \sup_{R\subseteq\mathcal{X}} |\Pr_{\mathcal{D}_\mathcal{X}}(R) - \Pr_{\mathcal{D}'_\mathcal{X}}(R)|$, then the following holds.

Observation 4.5. *Given two distributions $\mathcal{D}_\mathcal{X}$ and $\mathcal{D}'_\mathcal{X}$ such that TV$(\mathcal{D}_\mathcal{X}, \mathcal{D}'_\mathcal{X}) \leq \Lambda$, if a set of s points is an ϵ-net for $\mathcal{D}_\mathcal{X}$, then it is a $(\epsilon + \Lambda)$-net for $\mathcal{D}'_\mathcal{X}$.*

As a consequence of Observation 4.5 (proof in Appendix A) a classifier which is globally robust according to Theorem 4.4 for points sampled according to $\mathcal{D}_\mathcal{X}$ is also robust in $\mathcal{B}_{\delta,2\epsilon}$ for points sampled according to any $\mathcal{D}'_\mathcal{X}$, where $\mathrm{TV}(\mathcal{D}_\mathcal{X}, \mathcal{D}'_\mathcal{X}) \leq \epsilon$.

4.3. Global robustness guarantees through distillation

We can combine the arguments of the previous sections to devise a technique to infer the robustness of the teacher model with high probability. In our setting, we assume we can verify the global robustness of a student model $f_S(\mathbf{x})$ and are interested in inferring whether a teacher model $f_T(\mathbf{x})$ is globally robust or not. Given a (trained) teacher model, we obtain a student model through the gradient-aligned knowledge distillation discussed in Section 4.1. In particular, distillation is performed on a number of points s equivalent to the sample complexity provided by Theorem 4.4. If the student is verified to be globally 3δ-robust, then it is trivially 3δ-locally robust on the s points. According to Proposition 4.2 the teacher is then 3δ-locally robust on the s points. Using Theorem 4.4, we have that, with high probability, the teacher is also δ-globally robust in all the regions with high probability density.

Algorithm 1 summarizes the technique.

Algorithm 1 Check teacher robustness

```
1:  Input: teacher model $f_T : \mathbb{R}^m \to \mathbb{R}^n$, robustness
    radius $\delta$, parameters $\epsilon$, $\eta$, $\kappa$
2:  $s \leftarrow$ sample size given $\epsilon$ and $\eta$ (Theorem 4.4)
3:  $S \leftarrow$ sample $s$ points from $\mathcal{D}_\mathcal{X}$
4:  $f_S \leftarrow$ distill $f_T$ with a dataset of point-label pairs
    $(\mathbf{x}, f_T(\mathbf{x}))\ \forall \mathbf{x} \in S$ (Definition 4.1)
5:  if $f_S$ is $3\delta$-globally robust on $S$ then
6:      {$f_T$ is $3\delta$-locally robust on $S$ (Proposition 4.2)}
7:      return $f_T$ is $\delta$-globally robust with high probability
        (Theorem 4.4)
8:  else
9:      return $f_T$ not robust
10: end if
```

In line 3 in Algorithm 1 we assume we can (cheaply) sample unlabeled data from the data distribution $\mathcal{D}_\mathcal{X}$, which we can then label using f_T. In line 5 in Algorithm 1, global robustness (Definition 3.3) is defined with respect to the confidence threshold κ (Definition 3.2).

5. Experiments

We present preliminary empirical investigations on synthetic data. We consider a simple teacher model with tunable robustness properties constructed from a parameterized sigmoid function. We assess the quality of the distillation procedure under different conditions. We empirically show that robustness properties transfer from the teacher to the student after successful distillation. We compare the empirically assessed robustness of the student to the known robustness of the teacher. We vary the confidence threshold κ for robustness certification as well as the teachers gradients. This impacts both the teacher's robustness and the difficulty of distillation.

We use a simple binary decision problem using two truncated Gaussian distributions in $\mathbb{R}^2$, separated by a hyperplane. The teacher-model is constructed to perfectly classify this data using a parameterized sigmoid function with parameter $a \in (0, \infty)$,

$$f_T(\mathbf{x}, a) = \begin{pmatrix} \sigma(a\mathbf{x} \cdot \mathbf{1}) \\ 1 - \sigma(a\mathbf{x} \cdot \mathbf{1}) \end{pmatrix}. \tag{11}$$

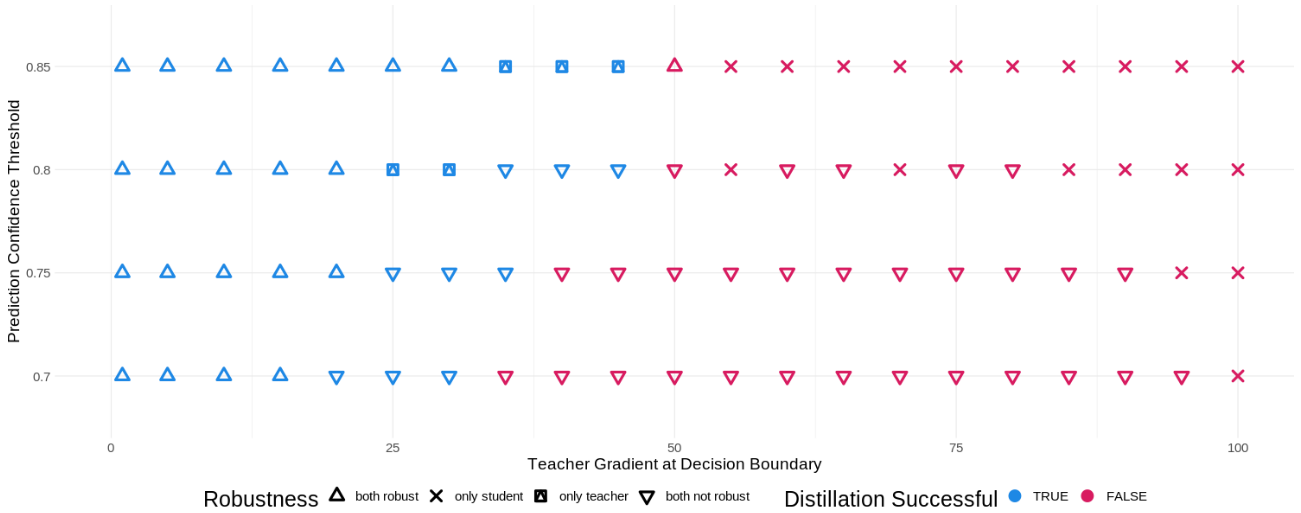

Figure 1. Robustness of student and teacher models on synthetic data experiment. Gradient of the teacher along the decision boundary was varied to control robustness properties. Coloring indicates whether the distillation was performed successfully.

The parameter a can be used to control the teacher's gradient

$$\nabla_{\mathbf{x}} f_T(\mathbf{x}, a) = \begin{pmatrix} \sigma(a\mathbf{x} \cdot \mathbf{1})(1 - \sigma(a\mathbf{x} \cdot \mathbf{1})) \\ -\sigma(a\mathbf{x} \cdot \mathbf{1})(1 - \sigma(a\mathbf{x} \cdot \mathbf{1})) \end{pmatrix}, \tag{12}$$

where $\mathbf{x} \cdot \mathbf{1}$ denotes the sum over the elements of $\mathbf{x}$. The parameter a allows us to control the teacher's gradients and robustness. To empirically assess the quality of the distillation we say that the distillation is successful if

$$\forall \mathbf{x} \in D_{\mathbf{x}} \quad \max(\text{softmax}(f_S(\mathbf{x})) \geq \max(\text{softmax}(f_T(\mathbf{x})), \tag{13}$$

that is, when the student is at least as confident for all predictions as the teacher. This prevents vacuous guarantees that do not hold for the teacher, where the student is defined to be robust only because it is non-confident.

The results of the experiments are presented in Figure 1. The shape of the data points conveys whether teacher and student agree on assessed robustness. The color indicates whether the distillation was successful as per the requirements in equation Equation (13), that is, whether the student is more confident than the teacher. We empirically observe that in all the cases where student and teacher do not agree on robustness, the student is less robust than the teacher. Hence, our procedure is tailored towards providing conservative robustness guarantees and a teacher is not reported to be robust when it is not. All the cases where we find the student to be robust whereas the teacher is not (denoted in Figure 1 with the symbol ×) are easily identified by our distillation quality measure Equation (13).

Our experiments empirically support our results and indicate that we can often transfer robustness properties from teacher to student model in real world settings, even if perfect distillation is not obtainable. In all cases where the robustness properties were not preserved after successful distillation, the student showed to be less robust, which indicates that conservative guarantees can still be given.

6. Conclusions

We have provided initial theoretical results for the use of gradient-aligned distillation to obtain a small student NN whose robustness can be easily certified using conventional formal verification methods. We derived a sample complexity bound for the distillation procedure that allows to certify that a teacher NN is globally robust with high probability if a student NN distilled from it is globally robust. Initial empirical results showed that gradient-aligned distillation indeed preserves robustness in practice and is useful to obtain conservative formal guarantees on the robustness of the teacher NN. In future work, we aim to integrate a formal verification tool like the one proposed by Athavale et al. (2024) to get robustness guarantees on real-world NNs. Furthermore, in cases where the student NN is not robust, we aim to investigate counter-examples that show precisely what leads to non-robustness. Finally, we will consider other learning theoretic techniques that may provide stronger robustness guarantees with a smaller sample complexity.

Acknowledgments and Impact Statement

This work was funded in part by the TU Wien DK SecInt; by the Vienna Science and Technology Fund (WWTF), project StruDL (ICT22-059); by the Austrian Science Fund (FWF), project NanOX-ML (6728); and by the by the European Unions Horizon Europe Doctoral Network programme under the Marie-Skłodowska-Curie grant, project Training Alliance for Computational systems chemistry (101072930).

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., and Havel, J. Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, 11(2):47–58, 2013. ISSN 1214-021X. doi: https://doi.org/10.2478/v10136-012-0031-x. URL https://www.sciencedirect.com/science/article/pii/S1214021X14600570.

Arora, R., Basu, A., Mianjy, P., and Mukherjee, A. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1J_rgWRW.

Athavale, A., Bartocci, E., Christakis, M., Maffei, M., Nickovic, D., and Weissenbacher, G. Verifying global two-safety properties in neural networks with confidence, 2024. URL https://arxiv.org/abs/2405.14400. To appear in CAV 2024.

Buciluǎ, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 535–541, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933395. doi: 10.1145/1150402.1150464. URL https://doi.org/10.1145/1150402.1150464.

Casadio, M., Komendantskaya, E., Daggitt, M. L., Kokke, W., Katz, G., Amir, G., and Refaeli, I. Neural network robustness as a verification property: A principled case study. In Shoham, S. and Vizel, Y. (eds.), *Computer Aided Verification*, pp. 219–231, Cham, 2022. Springer International Publishing. ISBN 978-3-031-13185-1. URL https://rdcu.be/dM7zG.

Chan, A., Tay, Y., and Ong, Y.-S. What It Thinks Is Important Is Important: Robustness Transfers Through Input Gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chan_What_It_Thinks_Is_Important_Is_Important_Robustness_Transfers_Through_CVPR_2020_paper.html.

Chen, Y., Wang, S., Qin, Y., Liao, X., Jana, S., and Wagner, D. Learning security classifiers with verified global robustness properties. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pp. 477–494, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384544. doi: 10.1145/3460120.3484776. URL https://doi.org/10.1145/3460120.3484776.

Croce, F., Andriushchenko, M., and Hein, M. Provable robustness of relu networks via maximization of linear regions. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2057–2066. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/croce19a.html.

Dudley, R. Balls in Rk do not cut all subsets of k + 2 points. *Advances in Mathematics*, 31 (3):306–308, 1979. ISSN 0001-8708. doi: https://doi.org/10.1016/0001-8708(79)90047-1. URL https://www.sciencedirect.com/science/article/pii/0001870879900471.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Gopinath, D., Katz, G., Pǎsǎreanu, C. S., and Barrett, C. Deepsafe: A data-driven approach for assessing robustness of neural networks. In Lahiri, S. K. and Wang, C. (eds.), *Automated Technology for Verification and Analysis*, pp. 3–19, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01090-4. URL https://rdcu.be/dM7HR.

Haussler, D. and Warmuth, M. The probably approximately correct (PAC) and other learning models. *The Mathematics of Generalization*, pp. 17–36, 2018. URL https://www.taylorfrancis.com/chapters/edit/10.1201/9780429492525-3.

Haussler, D. and Welzl, E. Epsilon-nets and simplex range queries. In *Proceedings of the Second Annual Sympo-*

sium on Computational Geometry, SCG '86, pp. 61–71, New York, NY, USA, 1986. Association for Computing Machinery. ISBN 0897911946. doi: 10.1145/10515.10522. URL https://doi.org/10.1145/10515.10522.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL http://arxiv.org/abs/1503.02531.

Huang, B., Chen, M., Wang, Y., Lu, J., Cheng, M., and Wang, W. Boosting accuracy and robustness of student models via adaptive adversarial distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24668–24677, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Huang_Boosting_Accuracy_and_Robustness_of_Student_Models_via_Adaptive_Adversarial_CVPR_2023_paper.html.

Huang, X., Kwiatkowska, M., Wang, S., and Wu, M. Safety verification of deep neural networks. In Majumdar, R. and Kunčak, V. (eds.), *Computer Aided Verification*, pp. 3–29, Cham, 2017. Springer International Publishing. ISBN 978-3-319-63387-9. URL https://rdcu.be/dM7KK.

Katz, G., Barrett, C., Dill, D. L., Julian, K., and Kochenderfer, M. J. Reluplex: An efficient smt solver for verifying deep neural networks. In Majumdar, R. and Kunčak, V. (eds.), *Computer Aided Verification*, pp. 97–117, Cham, 2017. Springer International Publishing. ISBN 978-3-319-63387-9. URL https://rdcu.be/dM7MB.

Katz, G., Huang, D. A., Ibeling, D., Julian, K., Lazarus, C., Lim, R., Shah, P., Thakoor, S., Wu, H., Zeljić, A., Dill, D. L., Kochenderfer, M. J., and Barrett, C. The marabou framework for verification and analysis of deep neural networks. In Dillig, I. and Tasiran, S. (eds.), *Computer Aided Verification*, pp. 443–452, Cham, 2019. Springer International Publishing. ISBN 978-3-030-25540-4. URL https://rdcu.be/dM7M0.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018. URL https://www.taylorfrancis.com/chapters/edit/10.1201/9781351251389-8.

Lee, G.-H., Alvarez-Melis, D., and Jaakkola, T. S. Towards Robust, Locally Linear Deep Networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SylCrnCcFX.

Lee, H., Cho, S., and Kim, C. Indirect gradient matching for adversarial robust distillation, 2023. URL https://arxiv.org/abs/2312.03286.

Leino, K., Wang, Z., and Fredrikson, M. Globally-robust neural networks. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6212–6222. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/leino21a.html.

Li, L., Xie, T., and Li, B. SoK: Certified Robustness for Deep Neural Networks. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1289–1310, 2023. doi: 10.1109/SP46215.2023.10179303. URL https://doi.org/10.1109/SP46215.2023.10179303.

Meng, M. H., Bai, G., Teo, S. G., Hou, Z., Xiao, Y., Lin, Y., and Dong, J. S. Adversarial robustness of deep neural networks: A survey from a formal verification perspective. *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2022. doi: 10.1109/TDSC.2022.3179131. URL https://doi.org/10.1109/TDSC.2022.3179131.

Mitzenmacher, M. and Upfal, E. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, USA, 2nd edition, 2017. ISBN 110715488X. URL https://dl.acm.org/doi/abs/10.5555/3134214.

Mustafa, N. H. and Varadarajan, K. Epsilon-approximations & epsilon-nets. In *Handbook of Discrete and Computational Geometry*, pp. 1241–1267. Chapman and Hall/CRC, 2017. URL https://www.taylorfrancis.com/chapters/edit/10.1201/9781315119601-47.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597, 2016. doi: 10.1109/SP.2016.41. URL https://doi.org/10.1109/SP.2016.41.

Rao, Q. and Frtunikj, J. Deep learning for self-driving cars: chances and challenges. In *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems*, SEFAIS '18, pp. 35–38, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357395. doi: 10.1145/3194085.3194087. URL https://doi.org/10.1145/3194085.3194087.

Sattelberg, B., Cavalieri, R., Kirby, M., Peterson, C., and Beveridge, R. Locally linear attributes of ReLU neural networks. *Frontiers in Artificial Intelligence*, 6, 2023. ISSN 2624-8212. doi: 10.3389/frai.2023.1255192. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1255192.

Seshia, S. A., Desai, A., Dreossi, T., Fremont, D. J., Ghosh, S., Kim, E., Shivakumar, S., Vazquez-Chanlatte, M., and Yue, X. Formal specification for deep neural networks. In Lahiri, S. K. and Wang, C. (eds.), *Automated Technology for Verification and Analysis*, pp. 20–34, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01090-4. URL https://rdcu.be/dM75I.

Shao, R., Yi, J., Chen, P.-Y., and Hsieh, C.-J. How and when adversarial robustness transfers in knowledge distillation? *arXiv preprint arXiv:2110.12072*, 2021.

Vaishnavi, P., Krish, V., Ahmed, F., Eykholt, K., and Rahmati, A. On the feasibility of compressing certifiably robust neural networks. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022. URL https://openreview.net/forum?id=YzPaQcK2Ko4.

Valiant, L. G. A theory of the learnable. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, STOC '84, pp. 436–445, New York, NY, USA, 1984. Association for Computing Machinery. ISBN 0897911334. doi: 10.1145/800057.808710. URL https://doi.org/10.1145/800057.808710.

Wang, L. and Yoon, K.-J. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3048–3068, 2022. doi: 10.1109/TPAMI.2021.3055564. URL https://doi.org/10.1109/TPAMI.2021.3055564.

Webb, S., Rainforth, T., Teh, Y. W., and Kumar, M. P. Statistical verification of neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1xcx3C5FX.

Wu, M., Wicker, M., Ruan, W., Huang, X., and Kwiatkowska, M. A game-based approximate verification of deep neural networks with provable guarantees. *Theoretical Computer Science*, 807:298–329, 2020. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2019.05.046. URL https://www.sciencedirect.com/science/article/pii/S0304397519304426. In memory of Maurice Nivat, a founding father of Theoretical Computer Science - Part II.

A. Proofs

A.1. Proof of Proposition 4.2

Proposition 4.2 (Robustness of the teacher model)**.** *Consider a teacher model f_T and a perfect student model f_S obtained through the distillation process described in Definition 4.1. $\forall \mathbf{x} \in D_{\mathbf{x}}$, if $f_S(\mathbf{x})$ is locally δ-robust in $\mathbf{x}$, then $f_T(\mathbf{x})$ is locally δ-robust in $\mathbf{x}$.*

Proof. Shao et al. (2021) proved the other direction of the implication in the proposition, with an analogous set of assumptions. Following a similar approach, the perfect student assumption for gradient-aligned distillation correspond to the conditions:

$$\begin{cases} f_S(\mathbf{x}) = f_T(\mathbf{x}) & \forall\, \mathbf{x} \in D_{\mathbf{x}}, \quad (14a) \\ \nabla_{\mathbf{x}} \mathcal{L}_{\text{CE}}(f_S(\mathbf{x}), y) = \nabla_{\mathbf{x}} \mathcal{L}_{\text{CE}}(f_T(\mathbf{x}), y) & \forall\, (\mathbf{x}, y) \in D. \quad (14b) \end{cases}$$

Following Shao et al. (2021), Equation (14a) implies that $\nabla_{\mathbf{x}} f_S(\mathbf{x}) = \nabla_{\mathbf{x}} f_T(\mathbf{x})$.

As any $\mathbf{x}' \in B_\delta(\mathbf{x})$ can be written as $\mathbf{x}' = \mathbf{x} + \delta'$ for $\|\delta'\| \le \delta$, $\forall\, \mathbf{x} \in D_{\mathbf{x}}$ and $\forall\, \mathbf{x}' \in B_\delta(\mathbf{x})$ it holds that:

$$\begin{aligned} f_T(\mathbf{x}') &= f_T(\mathbf{x} + \delta') \\ &= f_T(\mathbf{x}) + \delta' \nabla_{\mathbf{x}} f_T(\mathbf{x}) && \text{(local linearity)} \\ &= f_S(\mathbf{x}) + \delta' \nabla_{\mathbf{x}} f_S(\mathbf{x}) && \text{(Equations 14a, 14b)} \\ &= f_S(\mathbf{x} + \delta') = f_S(\mathbf{x}') \end{aligned}$$

Hence we have that:

$$f_T(\mathbf{x}') = f_S(\mathbf{x}'). \tag{15}$$

Now, note that:

$$\begin{aligned} \arg\max f_T(\mathbf{x}) &= \arg\max f_S(\mathbf{x}) && \text{(perfect student)} \\ &= \arg\max f_S(\mathbf{x}') && \text{(student } \delta\text{-robustness)} \\ &= \arg\max f_T(\mathbf{x}'). && \text{(using Equation (15))} \end{aligned}$$

That is, f_T is locally δ-robust in the sense of Definition 3.1. □

A.2. Proof of Proposition 4.3

Proposition 4.3 (Imperfect gradient alignment)**.** *Consider a teacher model $f_T(\mathbf{x})$ and a student model $f_S(\mathbf{x})$ such that $\forall\, (\mathbf{x}, y) \in D$ $f_S(\mathbf{x}) = f_T(\mathbf{x}) = y$. If $f_S(\mathbf{x})$ is locally δ-robust in $\mathbf{x}$, then $f_T(\mathbf{x})$ is locally δ_T-robust in $\mathbf{x}$, where*

$$\delta_T = \|\nabla_{\mathbf{x}} f_S(\mathbf{x})\| \|\nabla_{\mathbf{x}} f_T(\mathbf{x})\|^{-1} \delta. \tag{8}$$

Proof. Consider a distillation process performed, e.g., using a loss function analogous to that in Equation (3), where the student model provides the same predictions as the teacher, but where student and teacher models are not gradient-aligned. That is, assume $f_S(\mathbf{x}) = f_T(\mathbf{x}) = y\ \forall\, (\mathbf{x}, y) \in D$ but $\exists (\mathbf{x}, y) \in D \mid \nabla_{\mathbf{x}} f_S(\mathbf{x}) \neq \nabla_{\mathbf{x}} f_T(\mathbf{x})$. Let ρ be the ratio of the student and teacher gradient norms, i.e., $\rho = \|\nabla_{\mathbf{x}} f_S(\mathbf{x})\| \|\nabla_{\mathbf{x}} f_T(\mathbf{x})\|^{-1}$ at $\mathbf{x}$. Let f_S be δ-locally robust and let $\delta_T = \rho\delta$. We focus here on the case $\rho < 1$ which addresses the case of a teacher which is *less* robust than the student. Since any $\mathbf{x}' \in B_{\delta_T}(\mathbf{x})$, $\mathbf{x}'$ can be written as $\mathbf{x}' = \mathbf{x} + \delta'$, for $\delta' \le \delta_T \le \delta$, $\forall\, \mathbf{x} \in D_{\mathbf{x}}$ and $\forall\, \mathbf{x}' \in B_{\delta_T}(\mathbf{x})$, it holds that:

$$\arg\max f_S(\mathbf{x}') = \arg\max f_S(\mathbf{x}). \qquad \text{(local } \delta\text{-robustness)} \tag{16}$$

Now let $\delta'' = \rho\delta'$ and $\mathbf{x}'' = \mathbf{x} + \delta''$. Hence we have that:

$$\begin{aligned}
f_T(\mathbf{x}'') &= f_T(\mathbf{x}) + \delta''\nabla_{\mathbf{x}} f_T(\mathbf{x}) && \text{(local linearity)} \quad (17)\\
&= f_S(\mathbf{x}) + \delta''\nabla_{\mathbf{x}} f_T(\mathbf{x}) && \text{(perfect student)} \quad (18)\\
&= f_S(\mathbf{x}) + \rho\delta'\nabla_{\mathbf{x}} f_T(\mathbf{x}) && (19)\\
&= f_S(\mathbf{x}) + \frac{\|\nabla_{\mathbf{x}} f_S(\mathbf{x})\|}{\|\nabla_{\mathbf{x}} f_T(\mathbf{x})\|}\|\nabla_{\mathbf{x}} f_T(\mathbf{x})\|\,\|\delta'\|\cos\theta && \text{(dot product)} \quad (20)\\
&= f_S(\mathbf{x}) + \|\nabla_{\mathbf{x}} f_S(\mathbf{x})\|\,\|\delta'\|\cos\theta && (21)\\
&= f_S(\mathbf{x}) + \delta'\nabla_{\mathbf{x}} f_S(\mathbf{x}) && \text{(dot product)} \quad (22)\\
&= f_S(\mathbf{x}'). && \text{(local linearity)} \quad (23)
\end{aligned}$$

Finally,

$$\begin{aligned}
\arg\max f_T(\mathbf{x}) &= \arg\max f_S(\mathbf{x}) && \text{(perfect student)}\\
&= \arg\max f_S(\mathbf{x}') && \text{(student } \delta\text{-robustness)}\\
&= \arg\max f_T(\mathbf{x}''). && \text{(using equation (23))}
\end{aligned}$$

which allows us to conclude that f_T is δ_T-locally robust in $\mathbf{x}$. □

A.3. Proof of Theorem 4.4

Theorem 4.4. *Let ϵ, $\eta \in [0,1]$, $\delta \in \mathbb{R}_{>0}$. If a classifier $f(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}^n$ is locally 3δ-robust in*

$$s \geq \frac{8(m+1)}{\epsilon}\ln\frac{16(m+1)}{\epsilon} + \frac{4}{\epsilon}\ln\frac{2}{\eta} \tag{10}$$

points sampled from $\mathcal{D}_{\mathcal{X}}$, then it is globally δ-robust in any point covered by a ball in $\mathcal{B}_{\delta,\epsilon}$ with probability at least $1-\eta$.

Proof. Let N be a set of points with cardinality s, randomly sampled i.i.d. from $\mathcal{D}_{\mathcal{X}}$. Let us have a range space $(\mathcal{X}, \mathcal{B}_\delta)$, where $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{B}_\delta$ represents the set of δ-balls in $\mathcal{X}$. As per Theorem 3.8, N is an ϵ-net with high probability. Assuming N is an ϵ-net, it intersects each ball in $\mathcal{B}_{\delta,\epsilon}$. Now assume that f is 3δ-locally robust around each point in N. Let $B_\delta(\mathbf{x}_b) \in \mathcal{B}_{\delta,\epsilon}$ be the ball around some point $\mathbf{x}_b$ and let $\mathbf{x}_i \in N \cap B_\delta(\mathbf{x}_b)$. Let $B_{3\delta}(\mathbf{x}_i)$ represent the locally 3δ-robust ball around $\mathbf{x}_i$. Note that any ball $B_\delta(\mathbf{x}_b) \in \mathcal{B}_{\delta,\epsilon}$ is completely inside $B_{3\delta}(\mathbf{x}_i)$. Hence, if $B_{3\delta}(\mathbf{x}_i)$ is a robust ball, then so is $B_\delta(\mathbf{x}_b)$. Furthermore, since $B_\delta(\mathbf{x}_b)$ is also intersected by $\mathbf{x}_i$, the farthest point from $\mathbf{x}_i$ in $B_\delta(\mathbf{x}_b)$, say $\mathbf{x}'_b$, is at most 2δ away. Hence, a δ-ball around $\mathbf{x}'_b \in B_\delta(\mathbf{x}_b)$ is also completely contained in $B_{3\delta}(\mathbf{x}_i)$ and hence is robust. Consequently, any point in $\mathcal{B}_{\delta,\epsilon}$ is δ-robust.

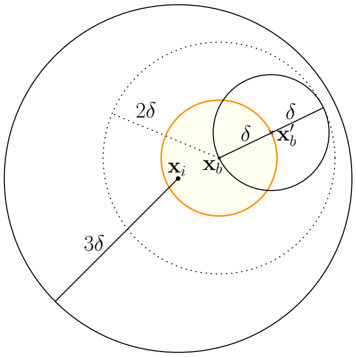

Figure 2. Visualization of the 3δ argument.

A.4. Proof of Observation 4.5

Observation 4.5. *Given two distributions $\mathcal{D}_{\mathcal{X}}$ and $\mathcal{D}'_{\mathcal{X}}$ such that TV$(\mathcal{D}_{\mathcal{X}}, \mathcal{D}'_{\mathcal{X}}) \leq \Lambda$, if a set of s points is an ϵ-net for $\mathcal{D}_{\mathcal{X}}$, then it is a $(\epsilon + \Lambda)$-net for $\mathcal{D}'_{\mathcal{X}}$.*

Proof. By the bound on TV$(\mathcal{D}_{\mathcal{X}}, \mathcal{D}'_{\mathcal{X}})$ we know $\forall R \in \mathcal{X} : |Pr_{\mathcal{D}_{\mathcal{X}}}(R) - \Pr_{\mathcal{D}_{\mathcal{X}}}(R)| \leq \Lambda$, especially $Pr_{\mathcal{D}'_{\mathcal{X}}}(R) \geq \epsilon + \Lambda$ implies $Pr_{\mathcal{D}'_{\mathcal{X}}}(R) \geq \epsilon$.

As the set of points intersects all $R \in \mathcal{X}$ s.t. $\Pr_{\mathcal{D}_{\mathcal{X}}}(R) \geq \epsilon$, we can conclude that it will also intersect all ranges $R \in \mathcal{X}$ s.t. $\Pr_{\mathcal{D}'_{\mathcal{X}}}(R) \geq \epsilon + \Lambda$ □