Certified Robustness against Sparse Adversarial Perturbations via Data Localization

Anonymous authors
Paper under double-blind review

Abstract

Recent work in adversarial robustness suggests that natural data distributions are localized, *i.e.*, they place high probability in small volume regions of the input space, and that this property can be utilized for designing classifiers with improved robustness guarantees for ℓ_2-bounded perturbations. Yet, it is still unclear if this observation holds true for more general metrics. In this work, we extend this theory to ℓ_0-bounded adversarial perturbations, where the attacker can modify a few pixels of the image but is unrestricted in the magnitude of perturbation, and we show necessary and sufficient conditions for the existence of ℓ_0-robust classifiers. Theoretical certification approaches in this regime essentially employ voting over a large ensemble of classifiers. Such procedures are combinatorial and expensive or require complicated certification techniques. In contrast, a simple classifier emerges from our theory, dubbed Box-NN, which naturally incorporates the geometry of the problem and improves upon the current state-of-the-art in certified robustness against sparse attacks for the MNIST and Fashion-MNIST datasets.

1 Introduction

It is by now well known that adversarial attacks affect Machine Learning (ML) systems that can potentially be used for security sensitive applications. However, despite significant efforts on robustifying ML models against adversarial attacks, it has been observed that their performance on most tasks under adversarial perturbation is not close to human levels. This motivated researchers to obtain theoretical impossiblity results for adversarial robustness Shafahi et al. (2018); Dohmatob (2019); Dai & Gifford (2022), which state that for general data distributions, no robust classifier exists against adversarial perturbations, even when the adversary is limited to making small ℓ_p-norm-bounded perturbations. However, such results are seemingly in conflict with the fact that humans can classify most natural images quite well under small ℓ_p-norm-bounded perturbations. Even more, there is a rich literature on certified robustness, *e.g.*, Zhang et al. (2018); Cohen et al. (2019); Pal & Vidal (2020); Fischer et al. (2020); Jeong & Shin (2020); Jia et al. (2022); Pfrommer et al. (2023); Salman et al. (2022); Eiras et al. (2022); Pal & Sulam (2023), where the goal is to obtain and analyze methods with provable guarantees on their robustness under adversarial attacks.

Pal et al. (2023) recently provided a solution to this apparent conflict, noting that existing impossibility results become vacuous when the data distribution is such that a large probability mass is concentrated on very small volume in the input space, a property they call (C, ϵ, δ)-concentration. This characterization implies that at least $1 - \delta$ probability mass is found in a region of volume at most $Ce^{-n\epsilon}$ for small $\delta \approx 0$ and large ϵ. As an example, this property dictates that sampling a random 224×224 dimensional image is extremely likely to *not* be a natural image. This property is intuitively satisfied for natural datasets like ImageNet, and Pal et al. (2023) formally show that whenever a classifier robust against small ℓ_2-bounded attacks exists for a data distribution (e.g., humans for natural images), this distribution must be concentrated. This shows that indeed, robust classifiers against ℓ_2 attacks can be obtained for natural image distributions, and there is no impossibility.

While these results are encouraging, attacks that are bounded in Euclidean norm have nice analytical properties that facilitated the results in Pal et al. (2023). In this work, we seek to understand if similar

notions can provide insights on provable defenses against *sparse* adversarial attacks (bounded in their ℓ_0 distance) where the adversary is limited to modifying a few pixels on the image, but those pixels can be modified in an unbounded fashion. Even though for humans it seems trivial to correctly classify a natural image corrupted in a few pixels, this problem has stood out as a particularly hard task for machine learning models. The difference is extreme: Su et al. (2019) demonstrated that adversarially modifying a *single* pixel leads to large performance degradation of many state of the art image recognition models. Standard ideas for improving robustness, like adversarial training, seem to be empirically ineffective against sparse attacks. Since then, researchers have resorted to enumerating a large number of subsets of the input pixels, and taking a majority vote over the class predicted from each subset, as a means of obtaining classifiers robust to sparse attacks. The resultant methods (Levine & Feizi, 2020b) are expensive, and need probabilistic certificates due to the combinatorial blow-up in the number of subsets needed as the number of attacked pixels increases. Follow-up work by Jia et al. (2022) has employed complicated certification schemes to reduce the slack in these certificates, while still remaining computationally expensive. Most recently, Hammoudeh & Lowd (2023) carefully selected these subsets to speed up the certificate computation. However, none of these existing methods utilizes the geometry of the underlying data distribution highlighted by our results. Departing from this stream of research, we propose a classifier that closely utilizes this underlying geometry to obtain robustness certificates. As a result, we provide a classifier that is lighter and simpler than all existing works, and an associated certification algorithm with ℓ_0 certificates that are better than prior work.

Our proof techniques extend results in Pal et al. (2023) to sparse adversarial attacks. In practice, one can always project the pixel values to lie in some predefined range, say $[0, 1]$, before classification, so we can consider adversarial perturbations to lie within $[0, 1]^n$ without any loss of generality. In other words, our adversary at power ϵ is allowed to modify an image from x to x' such that $\|x - x'\|_0 \leq \epsilon, \|x'\|_\infty \leq 1$. The techniques in Pal et al. (2023) break down under such an adversary, as their first assumption is to restrict attention to adversarial perturbations v such that $x + v$ cannot lie ϵ-close to the boundary of the image domain. In our case, the geometry of the problem is radically different: even a perturbation of size 1 is sufficient to take any image to the boundary of the domain $[0, 1]^n$ (simply perturb any pixel to 1). As a result, although we are motivated by Pal et al. (2023), our theory and certification algorithms are markedly different from those in that work.

In the above setting, we show that whenever there exists a classifier robust to adversarial modification of a few entries in the input, the underlying data distribution places a large mass, i.e., *localizes*, on low-volume subsets of the input space. We further show that the converse holds too, albeit with a strengthening of the localization condition; i.e., we show that when the data distribution localizes on low-volume subsets of the input space, and these subsets are sufficiently separated from one another, then a robust classifier exists. These results suggest that such underlying geometry in natural image distributions should be exploited for constructing classifiers robust against ℓ_0 attacks. Indeed, we then propose a simple classifier, called Box Nearest Neighbors (Box-NN), that utilizes this underlying geometry by having decision regions that are unions of axis-aligned rectangular boxes in the input space. Such a classifier naturally allows for ℓ_0 robustness certificates that improve upon prior work for certified defenses in a wide regime.

To summarize, we make the following contributions in this work:

1. In Section 2 we show that if a data-distribution p defining a multi-class classification problem admits a robust classifier whose error is at most δ under sparse adversarial perturbations to ϵ pixels, then there is a subset S of volume at most $Ce^{-\epsilon^2/n}$ and a class k such that the class conditional q_k places a large mass $q_k(S) \geq 1 - \delta$ on S, *i.e.*, q_k is $(C, \epsilon^2/n, \delta)$-localized.

2. In Section 3, we show that a stronger notion of localization, which ensures that the class conditional distributions are sufficiently separated with respect to the ℓ_0 distance, is sufficient for the existence of a robust classifier. In fact, this result generalizes to any distance d, showing the existence of a robust classifier w.r.t. perturbations bounded in distance d whenever the data distribution p is strongly localized with respect to d.

3. In Section 4, we propose a classifier certifiably robust against sparse adversarial attacks, called Box-NN, and derive certificates of ℓ_0 robustness for it. We then provide empirical evaluation on the

MNIST and the Fashion-MNIST datasets, and demonstrate that Box-NN obtains state-of-the-art results in certified ℓ_0 robustness.

2 Existence of an ℓ_0-Robust Classifier implies Localization

We will take our data domain to be $[0,1]^n$, to mimic the standard natural image classification tasks[1], *i.e.*, $\mathcal{X} = \{x \colon \|x\|_\infty \leq 1\}$. We will take our label domain to be $\mathcal{Y} = \{1, 2, \ldots, K\}$, and assume that we have a classification task defined by a data distribution p over $\mathcal{X} \times \mathcal{Y}$. The conditional distribution $p_{X|Y=k}$ for each $k \in \mathcal{Y}$ will be denoted by q_k.

For any classifier $f \colon \mathcal{X} \to \mathcal{Y}$, we recall the standard definition of robust risk $R_d(f, \epsilon)$ against perturbations bounded in a distance d as

$$R_d(f, \epsilon) = \mathop{\mathbb{P}}_{(x,y)\sim p} \left(\exists \bar{x} \in B_d(x, \epsilon) \text{ such that } f(\bar{x}) \neq y\right).$$

Similarly, we define a classifier f to be (ϵ, δ)-robust with respect to a distance d if the robust risk against perturbations at a distance bounded by ϵ is at most δ, *i.e.*, $R_d(f, \epsilon) \leq \delta$.

For the rest of this section, we will assume that p defines a task for which one can obtain a classifier f such that $R_{\ell_0}(f, \epsilon) \leq \delta$, where ϵ is a non-negative integer denoting the maximum number of pixels that an adversary can perturb. Given such an f, we will show that p should satisfy the special property of localization. In other words, we will obtain a necessary condition for ℓ_0 robustness. This special property of (C, ϵ, δ)-localization is similar to Pal et al. (2023, Definition 2.2), with a slight modification:

Definition 2.1 (Localized Distribution, modification of Pal et al. (2023))**.** A probability distribution q over a domain $\mathcal{X} \subseteq \mathbb{R}^n$ is said to be (C, ϵ, δ)-localized if there exists a subset $S \subseteq \mathcal{X}$ such that $q(S) \geq 1 - \delta$ but $\mathrm{Vol}(S) \leq C \exp(-\epsilon)$. Here, Vol denotes the standard Lebesgue measure on $\mathbb{R}^n$, and $q(S)$ denotes the measure of S under q.

Definition 2.1 is similar to Pal et al. (2023, Definition 2.2) but it removes the explicit dimension of the problem, *i.e.*, n, from the volume constraint. This allows one to state the results in Pal et al. (2023), as well as ours, under the same definition. Additionally, we rename the property from *concentration* in Pal et al. (2023) to *localization*, in order to distinguish ourselves from the well known notion of *concentration of measure*. These two notions are related and, before proceeding, we compare them in more detail.

The notion of *measure concentration* from high dimensional probability theory roughly states that for a given large dimension n, "a well behaved function h of the random variables $Z_1, Z_2, \ldots, Z_n$ takes values close to its mean $\mathbb{E}\, h(Z_1, \ldots, Z_n)$ with high probability" (Talagrand, 1996). A popular quantification of this notion states that for a metric space $(\mathcal{X}, d)$ and a probability distribution q over $\mathcal{X}$, the concentration function α defined as

$$\alpha_{q,d}(t) = \sup_{S \subseteq \mathcal{X},\ q(S) \geq 1/2} 1 - q(S^{+t}), \tag{1}$$

decreases "very fast" with t, where recall that $S^{+t} = \{x \in \mathcal{X} \colon d(x, S) \leq t\}$. We typically say that q has the property of measure concentration if there is an exponential decay as $\alpha_{q,d}(t) \sim \exp(-\gamma t)$ for all $t \geq 0$, and some universal constant γ.

In contrast, the definition of (C, ϵ, δ)-localization requires the existence of $S \subseteq \mathcal{X}$ such that $q(S) \geq 1 - \delta$ and $\mathrm{Vol}\,(S) \leq C \exp(-\epsilon)$. Concentration and localization are similar in the underlying message: most of the mass in q is concentrated near a small region in space. However, the mathematical formalization is different, as localization does not require a fast enough rate of decay of the measure, and hence does not require an underlying metric on the space $\mathcal{X}$. In order to show that a given distribution q localizes, it is sufficient to provide a single instance of a set $S \subseteq \mathcal{X}$ that satisfies the localization parameters. For our data domain $\mathcal{X} = [0,1]^n$, we will consider a family of probability distributions given by $q_a = \mathrm{Unif}([0,a]^n)$ for $a \in (0,1]$, and comment on their localization and measure concentration parameters, to shed light into their similarities and differences.

[1] Albeit with a scaling – natural images are typically stored with each pixel value in $[0, 255]$.

For any $S \subseteq [0,a]^n \subseteq \mathcal{X}$, we can simplify $1-\delta \leq q_a(S) = \frac{1}{a^n}\mathrm{Vol}\,(S) \leq \frac{1}{a^n}\exp(-\epsilon)$ to obtain that

$$q_a \text{ is } \left(1, \log\left(\frac{1}{1-\delta}\right) + n\log\left(\frac{1}{a}\right), \delta\right) - \text{localized for any } \delta \in [0,1].$$

From the above we can see that keeping $\delta, a < 1$ fixed, q_a becomes "more localized" as the dimension n increases. Similarly, keeping δ, n fixed, q_a becomes more localized as a gets closer to 0. In this sense, the localization parameters depend on the scale of the support of the underlying distribution.

In contrast, as measure concentration depends on an underlying metric, the concentration parameters are independent of the scale of the support when the metric is invariant to scaling. As an example, for $\mathcal{X}$ equipped with the hamming metric, $d_0(x,x') = \|x-x'\|_0$, the concentration function for the distribution q_a can be shown to be

$$\alpha_{q_a,d_0}(t) \leq 2\exp\left(-\frac{t^2}{n}\right). \tag{2}$$

Armed with the above definition, we will now derive a necessary condition for ℓ_0-robustness in terms of localization, by using a measure-concentration result w.r.t. the ℓ_0 distance due to Talagrand (1995).

Theorem 2.2. *If there exists an (ϵ,δ)-robust classifier f with respect to the ℓ_0 distance for a data distribution p, then at least one of the class conditionals $q_1, q_2, \ldots, q_K$ must be $(C, \epsilon^2/n, \delta)$–localized according to Definition 2.1. Further, if the classes are balanced, then all the class conditionals are $(C_{\max}, \epsilon^2/n, K\delta)$-localized. Here, C and $C_{\max}$ are constants dependent on f.*

Proof. We are given a classifier f which is (ϵ,δ)-robust w.r.t. perturbations bounded in the ℓ_0 distance. In other words, we have $R_{\ell_0}(f,s) \leq \delta$. Expanding this we get

$$\sum_k \mathbb{P}\left(\exists \bar{x} \in B_{\ell_0}(x,\epsilon) \text{ such that } f(\bar{x}) \neq k\right)\mathbb{P}(y=k) \leq \delta.$$

In other words, there exists a class k' satisfying $q_{k'}\big(\{x \in \mathcal{X} \colon \exists \bar{x} \in B_{\ell_0}(x,\epsilon) \text{ such that } f(\bar{x}) \neq k'\}\big) \leq \delta$. Defining the unsafe set for the class k' as $U_{k'} = \{x \in \mathcal{X} \colon \exists \bar{x} \in B_{\ell_0}(x,\epsilon) \text{ such that } f(\bar{x}) \neq k'\}$, we have shown

$$q_{k'}(U_{k'}) \leq \delta. \tag{3}$$

Define $A_{k'} \subseteq \mathcal{X}$ to be the region where f predicts k', *i.e.*, $A_{k'} = \{x \in \mathcal{X} \colon f(x) = k'\}$. Further, for any set Z define $Z^{+\epsilon}$ to be all the points in the domain $\mathcal{X}$ which are at most s away from Z in ℓ_0 distance, *i.e.*, $Z^{+\epsilon} = \{x \in \mathcal{X} \colon \exists \bar{x} \in Z \text{ such that } \|x-\bar{x}\|_0 \leq \epsilon\}$ Then, we have

$$\begin{aligned} U_{k'} &= \{x \in \mathcal{X} \colon \exists \bar{x} \text{ such that } \|x-\bar{x}\|_0 \leq \epsilon, f(\bar{x}) \neq k'\} \\ &= \{x \in \mathcal{X} \colon \exists \bar{x} \in (\mathcal{X} \setminus A_{k'}) \text{ such that } \|x-\bar{x}\|_0 \leq \epsilon\} \\ &= (\mathcal{X} \setminus A_{k'})^{+\epsilon}. \end{aligned}$$

Now, we will use measure concentration on the unit cube from Talagrand (1995, Proposition 2.1.1):

Lemma 2.3 (Proposition 2.1.1 in Talagrand (1995))**.** *For $B \subseteq [0,1]^n$, $\mathrm{dist}(x,B) = \min_{z\in B}\|x-z\|_0$, any measure μ on $[0,1]$, we have*

$$\mathop{\mathbb{P}}_{x\sim\mu^n}(\mathrm{dist}(B,x) \geq t) \leq \frac{1}{\mathbb{P}_{x\sim\mu^n}(x \in B)}\exp(-t^2/n).$$

Note that since the domain $[0,1]^n$ has n-dimensional volume 1, *i.e.*, $\mathrm{Vol}([0,1]^n) = 1$, the uniform measure of any set $\mu^n(B) = \mathrm{Vol}(B)$, for $B \subseteq [0,1]^n$. Substituting $B = \mathcal{X} \setminus A_{k'}$, $t = \epsilon$, $\mu = \mathrm{Unif}([0,1])$, in Lemma 2.3, we obtain

$$\mathrm{Vol}(\mathcal{X} \setminus A_{k'})^{+\epsilon} \geq 1 - \frac{\exp(-\epsilon^2/n)}{\mathrm{Vol}(\mathcal{X} \setminus A_{k'})}.$$

Using $\mathrm{Vol}(\mathcal{X} \setminus U_{k'}) = 1 - \mathrm{Vol}(\mathcal{X} \setminus A_{k'})^{+\epsilon}$, we obtain

$$\mathrm{Vol}(\mathcal{X} \setminus U_{k'}) \leq \frac{\exp(-\epsilon^2/n)}{\mathrm{Vol}(\mathcal{X} \setminus A_{k'})}. \tag{4}$$

Finally, combining (3), (4), and taking $S = \mathcal{X} \setminus U_{k'}$, we have

$$q_{k'}(S) \geq 1 - \delta, \qquad \mathrm{Vol}(S) \leq C \exp(-\epsilon^2/n),$$

where $C = \frac{1}{1-\mathrm{Vol}(A_{k'})}$, showing that $q_{k'}$ is $(C, \epsilon^2/n, \delta)$-localized. If the classes were balanced, repeating the above argument for each class shows that q_k is $(C, \epsilon^2/n, K\delta)$-localized for all $k \in \mathcal{Y}$ for $C_{\max} = \max_{k'}(1/(1 - \mathrm{Vol}(A_{k'})))$. □

Discussion on Theorem 2.2 A few comments are in order for the above result.

1. Theorem 2.2 demonstrates that whenever a ℓ_0 robust classifier exists for a data distribution, this distribution must be localized. This could be instantiated for real data sets like ImageNet to obtain interesting observations about the underlying distribution. For instance, humans are robust to perturbation of a few pixels to any image in ImageNet. Then, Theorem 2.2 tells us that ImageNet is localized. Note, however, that the localization *parameters* (*i.e.*, C, ϵ, δ for the human classifier) are unknown.

2. The localization parameters in Theorem 2.2 are different than the concentration parameters in Pal et al. (2023, Theorem 2.1). Specifically, Pal et al. (2023, Theorem 2.1) shows that $(C, n\epsilon, \delta)$-concentration is a necessary condition for ℓ_2-robustness under Definition 2.1, and we will now show that $(C, \epsilon^2/n, \delta)$-localization is a necessary condition for ℓ_0-robustness. This demonstrates that the existence of a classifier robust to ℓ_0 classifier implies a different kind of localization of the data distribution than robustness to ℓ_2 perturbations. While Pal et al. (2023) assume that their data lies in a unit ℓ_2 ball with adversarial perturbation strength $\epsilon \in [0, 1]$, we assume that our data lies in a unit ℓ_∞ ball and with perturbation strength $\epsilon \in \{0, 1, 2, \ldots, n\}$. As such a direct comparison of the parameters is not immediate as our work deals with objects very different from Pal et al. (2023).

3. Theorem 2.2 suggests that for obtaining ℓ_0 robust classifiers, we should try to find and classify over the sets that the distribution localizes on. This is a significant departure from the existing literature on ℓ_0-robust classifiers Levine & Feizi (2020a); Jia et al. (2022); Hammoudeh & Lowd (2023), and indeed, we will obtain a classifier in Section 4 that respects such geometry.

We have now demonstrated that localization is a necessary condition for the existence of a classifier robust to perturbations bounded in the ℓ_0 distance, *i.e.*, perturbations having a small support. Inspired by the investigations in Pal et al. (2023), we will now consider whether this condition is also sufficient.

3 d-Strong Localization implies Existence of a d-Robust Classifier

Localization of the data distribution ensures that each class conditional concentrates on a small volume subset of $\mathcal{X}$. However, as noted in Pal et al. (2023), these subsets might intersect too much, in which case there might not exist a classifier with low standard risk, *i.e.*, $R_{\ell_0}(f, 0)$. Hence, one cannot expect localization to be sufficient for the existence of a classifier with low robust risk, *i.e.*, $R_{\ell_0}(f, \epsilon)$ with $\epsilon > 0$. However, if these subsets were *separated* enough, then one can expect to use them to build a robust classifier. Indeed, we will now formalize this intuition to obtain a condition stronger than localization, which will be shown to be sufficient for the existence of a robust classifier.

Definition 3.1 (d-Strongly Localized Distributions, generalizing Pal et al. (2023))**.** A distribution p is said to be $(\epsilon, \delta, \gamma)$-strongly-localized with respect to a distance d, if each class conditional distribution q_k localizes over the set $S_k \subseteq \mathcal{X}$ such that $q_k(S_k) \geq 1 - \delta$, and $q_k\left(\bigcup_{k' \neq k} S_{k'}^{+2\epsilon}\right) \leq \gamma$, where $S^{+\epsilon}$ denotes the ϵ-expansion of the set S in d, *i.e.*, $S^{+\epsilon} = \{x \colon \exists \bar{x} \in S \text{ such that } d(x, \bar{x}) \leq \epsilon\}$.

With the above definition, we will now obtain a generalization of Pal et al. (2023, Theorem 3.1) to an arbitrary distance d:

Theorem 3.2. *If p is $(\epsilon, \delta, \gamma)$-strongly localized with respect to a distance d, then there exists a classifier f such that $R_d(f, \epsilon) \leq \delta + \gamma$.*

Proof. At a high level, we will construct a classifier g that predicts the label k over an ϵ-expansion of the set S_k on which the class conditional q_k localizes. We will then "shave off" some regions from each S_k to ensure g is well defined. For the rest of the input space $\mathcal{X}$ we will predict an arbitrary label, as we incur at most γ in robust risk. Our construction of the robust classifier f is same as that in Pal et al. (2023), extended to general d. However, bounding the robust risk of f needs technical innovations, since we are bounding the robust risk with respect to a general distance d, as opposed to the ℓ_2 norm in Pal et al. (2023).

For each $k \in \{1, 2, \ldots, K\}$, let S_k be the set over which the conditional density q_k is localized, i.e., $q_k(S_k) \leq 1-\delta$. Define $S^{+\epsilon}$ to be the ϵ-expansion of the set S, as $S^{+\epsilon} = \{x \colon \exists x' \in S, d(x, x') \leq \epsilon\}$. Define C_k to be the ϵ-expanded version of the localized region S_k but removing the ϵ-expanded version of all other regions $S_{k'}$, as

$$C_k = \left(S_k^{+\epsilon} \setminus \cup_{k' \neq k} S_{k'}^{+\epsilon}\right) \cap \mathcal{X}.$$

Similar to the construction in Pal et al. (2023), we will use these regions to define the classifier $f \colon \mathcal{X} \to \{1, 2, \ldots, K\}$ as

$$f(x) = \begin{cases} 1, & \text{if } x \in C_1 \\ 2, & \text{if } x \in C_2 \\ \vdots & \\ K, & \text{if } x \in C_K \\ 1, & \text{otherwise} \end{cases}.$$

We will now show that $R_d(f, \epsilon) \leq \delta + \gamma$, which can be recalled to be

$$R_d(f, \epsilon) = \sum_k q_k(U_k) p_Y(y = k), \tag{5}$$

where the q_k mass in (5) is over the set of all points $x \in \mathcal{X}$ that admit an ϵ-adversarial example for the class k, defined as

$$U_k = \{x \in \mathcal{X} \colon \exists \bar{x} \in B_d(x, \epsilon) \cap \mathcal{X} \text{ such that } f(\bar{x}) \neq k\}. \tag{6}$$

As we saw earlier in the proof of Theorem 2.2, $U_k = (\mathcal{X} \setminus C_k)^{+\epsilon} \cap \mathcal{X}$. We will obtain an upper bound on $q_k(U_k)$, which will in turn give us an upper bound on $R_d(f, \epsilon)$.

Let $A = S_k^{+\epsilon} \cap \mathcal{X}$ and $B = \cup_{k' \neq k} S_{k'}^{+\epsilon}$. As $C_k = A \setminus B$, we have

$$\begin{aligned} \mathcal{X} \setminus C_k &= \mathcal{X} \cap (A \cap B^c)^c \\ &= \mathcal{X} \cap (A^c \cup B) \\ &= (\mathcal{X} \cap A^c) \cup (\mathcal{X} \cap B) \\ &= \left(\mathcal{X} \cap \left(S_k^{+\epsilon}\right)^c\right) \cup \left(\cup_{k' \neq k} (\mathcal{X} \cap S_{k'}^{+\epsilon})\right). \end{aligned}$$

Then, we can expand $(\mathcal{X} \setminus C_k)^{+\epsilon}$

$$\left(\mathcal{X} \cap \left(S_k^{+\epsilon}\right)^c\right)^{+\epsilon} \cup \left(\cup_{k' \neq k} (\mathcal{X} \cap S_{k'}^{+\epsilon})^{+\epsilon}\right),$$

from the property $(U \cup V)^{+\epsilon} = U^{+\epsilon} \cup V^{+\epsilon}$. Now, since all the mass of q_k lies in $\mathcal{X}$, *i.e.*, $q_k(\mathcal{X}) = 1$, we have $q_k(\mathcal{X} \cap V) = q_k(V)$ for any set V. Applying this, we have

$$\begin{aligned} q_k(U_k) &= q_k(\mathcal{X} \setminus C_k)^{+\epsilon} \\ &\leq q_k\left(\mathcal{X} \cap \left(S_k^{+\epsilon}\right)^c\right)^{+\epsilon} + q_k\left(\cup_{k' \neq k} (\mathcal{X} \cap S_{k'}^{+\epsilon})^{+\epsilon}\right) \\ &\leq q_k\left(\left(S_k^{+\epsilon}\right)^c\right)^{+\epsilon} + q_k\left(\cup_{k' \neq k} (S_{k'}^{+\epsilon})^{+\epsilon}\right). \end{aligned}$$

Now applying Lemma A.1 we have $\left(\left(S_k^{+\epsilon}\right)^c\right)^{+\epsilon} = \left(\left(S_k^{+\epsilon}\right)^{-\epsilon}\right)^c$. Again from Lemma A.1 we know that $(V^{+\epsilon})^{-\epsilon} \supseteq V$ for any set V. Hence, we have $\left(\left(S_k^{+\epsilon}\right)^{-\epsilon}\right)^c \subseteq S_k^c$. Continuing,

$$\begin{aligned} q_k(U_k) &\leq q_k(S_k^c) + q_k\left(\cup_{k' \neq k} S_{k'}^{+2\epsilon}\right) \\ &\leq \delta + \gamma, \end{aligned}$$

Finally, as $\sum_k p_Y(y = k) = 1$, from (6) we have $R_d(f, \epsilon) \leq \delta + \gamma$. □

We note that (Pal et al., 2023, Theorem 3.2) follows as a direct corollary of our result Theorem 3.2 by taking d to be the ℓ_2 distance.

Implications for Existing Impossibility Results In our setting, Shafahi et al. (2018) prove that for any classifier $f\colon \mathcal{X} \to \{1, 2, \ldots, K\}$ for any class k with $P(Y = k) \leq 1/2$, any point $x \sim q_k$ is either mis-classified, or admits an ϵ-adversarial example with probability at least

$$1 - \beta_{q_k} \exp\left(-\epsilon^2/n\right), \tag{7}$$

where $\beta_{q_k} = 2 \sup_x q_k(x)$ depends on the class conditional q_k. When q_k is localized, β_{q_k} can grow faster than $\exp\left(-\epsilon^2/n\right)$, making the lower bound vacuous. This implies that *for localized data-distributions there is no impossibility, and there is a wide class of high-dimensional classification problems for which robust classifiers exist.* We now provide a concrete example.

Example 3.1. *Let us consider a problem with 2 classes defined by the distribution p such that $P(Y = 0) = P(Y = 1) = 1/2$, the class conditional $q_1 = P(X|Y = 1) = \mathrm{Unif}(B_{\ell_\infty}(\mathbf{1}, \epsilon))$, and similarly $q_2 = P(X|Y = 2) = \mathrm{Unif}(B_{\ell_\infty}(-\mathbf{1}, \epsilon))$. For this distribution, $\beta_{q_1} = \beta_{q_2} = \exp(n)$, and the lower bound (7) becomes vacuous for $\epsilon \leq \sqrt{n}$ as*

$$1 - \beta_{q_k} \exp\left(-\epsilon^2/n\right) = 1 - 2\exp(-\epsilon^2/n + n) \leq 0.$$

Even though Example 3.1 is quite simple, the construction of small ℓ_∞ balls in the input space containing most of the mass of the distribution is quite general, and depicts a wide class of data-distributions where existing impossibility results are vacuous. We will now demonstrate that these general theoretical ideas lead to practical ℓ_0 robust classifiers.

4 ℓ_0-Adversarially Robust Classification via the Box-NN classifier

In this section, our aim will be to derive a ℓ_0-robust classifier by utilizing the geometry exposed by Theorem 3.2. To this end, we will first investigate how a robust classifier looks like for a simple 2-class problem in 3-dimensions. This will motivate a general form of a classifier whose decision regions are axis-aligned cuboids, or boxes. Finally, we will generalize this classifier to obtain a ℓ_0-robust classifier and derive corresponding ℓ_0 certificates.

4.1 Development and Robustness Certification

Consider $n = 3$, and say there are two classes, CAT and DOG, defining conditional distributions q_1 and q_2, strongly localized over S_1 and S_2 respectively, such that $q_1(S_2^{+1}) = 0$ and $q_2(S_1^{+1}) = 0$. In such a situation, Theorem 3.2 (invoked with $\epsilon = 1$) constructs a robust classifier f_A as the following:

$$f_A(x) = \begin{cases} \text{dog}, & \text{if } x \in S_1^{+1} \\ \text{cat}, & \text{if } x \in S_2^{+1} \\ \text{cat}, & \text{otherwise.} \end{cases}.$$

However, in practice, S_1, S_2 might be very complex, and hence f_A might be computationally hard to evaluate. For instance, Fig. 1 shows an illustration where these sets (shaded green and orange) have complicated shapes.

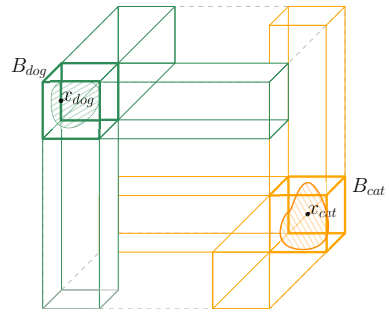

Figure 1: S_1 is the green shaded region around x_{dog}, where the class dog is localized, and S_2 is the orange shaded region around x_{cat}, where the class cat is localized.

From Fig. 1, we see that the classifier f_A is robust to 1-pixel perturbations whenever $x \in S_1$ or $x \in S_2$, as Theorem 3.2 predicts. More importantly, we see that a perturbation of a single pixel of any $x_{cat} \in S_2$ lies within the union of the orange cuboids. In other words, $\{x' \in [0,1]^3 \colon \|x - x\|_0 \leq 1, x \in S_1\} = S_1^{+1} \subseteq$ ORANGE, and similarly for the dog class. Furthermore, we see that the intersection of these orange cuboids is given by the cube B_{cat}. We can see that for any $x \in B_{cat}$, no single-pixel perturbation v can take $x + v$ outside the orange region ORANGE, and similarly for the dog class. However, B_{cat}, B_{dog} are very efficiently described, they are simply axis-aligned polyhedra enclosing S_2 and S_1 respectively. This motivates our modified classifier f_B,

$$f_B(x) = \begin{cases} \text{dog}, & \text{if } x \in B_{dog}^{+1} \\ \text{cat}, & \text{if } x \in B_{cat}^{+1} \\ \text{cat}, & \text{otherwise.} \end{cases}.$$

While f_B is efficient to describe, it ignores a large portion of the input region outside the green and the orange cuboids, *i.e.*, $\mathcal{X} \setminus B_{dog}^{+1} \cup B_{cat}^{+1}$, by making the constant prediction cat in this region. We can further extend f_B to attempt to correctly classify those regions as well, by computing ℓ_0 distances to our boxes B_{cat}, B_{dog}, as

$$f_C(x) = \underset{y \in \{\text{cat,dog}\}}{\arg\min} \ \text{dist}(x, B_y),$$

where

$$\text{dist}(x, S) = \min_v \|v\|_0 \text{ sub. to } x + v \in S \tag{8}$$

gives the minimum number of pixel changes needed to get from x to S. While solving (8) is computationally hard for general S, the following lemma shows that for our axis-aligned boxes B, (8) can be computed efficiently, in closed form. The proofs of all our results can be found in Appendix A.

Lemma 4.1 (ℓ_0 distance to axis-aligned boxes)**.** *For an axis aligned box $B(a,b)$ specified as $B(a,b) = \{x \colon a \leq x \leq b\}$, where $a, b, x \in \mathbb{R}^n$, and all inequalities are element-wise, we have*

$$\text{dist}(x, B(a,b)) = \sum_{i=1}^{n} \mathbf{1}\left(x_i \notin [a_i, b_i]\right),$$

which can be computed in $O(n)$ operations.

For real data distributions, however, having a single box per class would be overly simplistic and not provide good accuracy. Thus, we generalize f_C to our BOX-NN classifier operating on boxes $\mathcal{B} = \{B_1, B_2, \ldots, B_M\}$, such that we have an label $y_m \in \{1, 2, \ldots, K\}$ associated with each B_m. Our BOX-NN classifier is then defined as

$$\text{BOX-NN}(x, \mathcal{B}) = y_{m^\star}, \text{ where } m^\star = \underset{m}{\arg\min}\, \text{dist}(x, B_m).$$

Note that, so far, we have not described how these boxes $\mathcal{B}$ are learned from data. This will be the subject of Section 4.2 and onward. We can now obtain a ℓ_0 robustness certificate for BOX-NN via the following Theorem.

Theorem 4.2 (Robustness Certificate for Box-NN)**.** *Given a set of boxes $\mathcal{B}$ and their associated labels $\{y_m\}_{m=1}^M$, define*

$$m^\star = \arg\min_m \operatorname{dist}(x, B_m), \quad d_1 = \operatorname{dist}(x, B_{m^\star}),$$

and

$$d_2 = \min_{m\colon y_m \neq y_{m^*}} \operatorname{dist}(x, B_m).$$

Then, with $\operatorname{margin}(x) \stackrel{\text{def}}{=} d_2 - d_1$*, we have* $\text{Box-NN}(x, \mathcal{B}) = \text{Box-NN}(x', \mathcal{B})$ *whenever* $\|x' - x\|_0 < \operatorname{margin}(x)/2$.

Key Intuition Our robust classifier Box-NN is essentially a generalization of the nearest-neighbor classifier to a nearest-box classifier, specifically suited to ℓ_0 metrics. This simple form turns out to be the right choice, in the sense of the theoretical motivation of our previous section, for defending against sparse perturbations. As we will shortly see, Box-NN also empirically produces better certificates than prior work in several regimes.

Having developed the geometric intuition and the theoretical robustness guarantees for Box-NN, we will now describe how we learn our classifier from data, and the associated challenges.

4.2 Learning Box-NN from Data

In this section, we are concerned with learning boxes $\{B_m\}$ and their associated labels $\{y_m\}$, such that Box-NN obtains a high accuracy under sparse adversarial perturbations. For the rest of this section, we will refer to the classifier Box-NN as f_θ, with the learnable parameters $\theta = \{a_k, b_k, y_k\}_{k=1}^M$ following the notation in Lemma 4.1.

The quantity we are interested in maximizing is the robust accuracy, defined as $1 - R_{\ell_0}(f_\theta, \epsilon)$ following our notation in Section 2. As we do not have access to the data distribution, we will instead be concerned with maximizing the empirical robust accuracy $\operatorname{RobustAcc}(f_\theta, \epsilon)$ defined over a set of samples $\{x_i, y_i\}_{i=1}^N$ given by

$$\frac{1}{N}\sum_{i=1}^N \mathbf{1}\left[\forall x'\colon \|x' - x_i\|_0 \leq \epsilon,\ f_\theta(x') = y_i\right]. \tag{9}$$

The objective in (9) is a complicated object, and direct maximization w.r.t. θ is challenging. In the following, we will first lower bound (9) and then use several optimization tricks to efficiently maximize this lower bound.

Recall from Theorem 4.2 that $f_\theta(x) = f_\theta(x')$ for all x' satisfying $\|x - x'\|_0 \leq C_\theta(x) \stackrel{\text{def}}{=} \operatorname{margin}(x)/2$, where C_θ is a pointwise certificate (at x) of robustness for f_θ. With this, we have the certified accuracy lower bound $\operatorname{RobustAcc}(f_\theta, \epsilon) \geq \operatorname{CertAcc}(f_\theta, \epsilon)$ defined as

$$\operatorname{CertAcc}(f_\theta, \epsilon) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^N \mathbf{1}[f_\theta(x_i) = y_i] \cdot \mathbf{1}[C_\theta(x_i) \geq \epsilon]. \tag{10}$$

We will take a gradient based optimization approach to maximize (10) over θ. However, since the gradients of $\mathbf{1}[\cdot]$ are zero almost everywhere (and discontinuous otherwise), we will progressively relax the indicators in (10). To this end, we maximize the integral of $\operatorname{CertAcc}(f_\theta, \epsilon)$ over all $\epsilon \geq 0$ instead of treating it point-wise[2], leading to the objective

$$L_1(\theta) = \frac{1}{N}\sum_{i=1}^N \mathbf{1}[f_\theta(x_i) = y_i] \cdot C_\theta(x_i). \tag{11}$$

On the other hand, recall from Theorem 4.2 that the margin involves the min function,

$$\operatorname{margin}(x) = \min_m \operatorname{dist}(x, B_m) - \min_{m\colon y_m \neq y_{m^\star}} \operatorname{dist}(x, B_m).$$

[2] *i.e.*, $\int_{\epsilon \geq 0} \mathbf{1}[\epsilon \leq \alpha] d\epsilon = \alpha$

The gradient of min w.r.t. its input $(c_1, \ldots, c_M)$ is extremely sparse[3], and hence a very small number of parameters θ_i are updated at each step of gradient descent using gradients of (11). As a result, optimization is extremely slow. We remedy this by using a soft approximation to min which has dense gradients,

$$\min_\tau\{c_1, \ldots, c_M\} \stackrel{\text{def}}{=} \sum_{m=1}^{M} c_m \frac{\exp(-\tau c_m)}{\sum_j \exp(-\tau c_j)}, \tag{12}$$

where τ is a parameter that approximately controls the sparsity of the gradients. The function $\min_\tau$ is equal to min in the limit $\tau \to \infty$, and reduces to the average when $\tau = 0$. This step is crucial for the performance of our method.

Furthermore, we find that for many data points x_i, a small number of boxes m contribute a lot to the final loss due to large distances $\text{dist}(x_i, B_m)$. As a result, learning is slow for parameters corresponding to the remaining boxes. To prevent such imbalance, we clip the certificates to 50. With these approximations, we obtain

$$L_2(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[f_\theta(x_i) = y_i] \cdot \tilde{C}_\theta(x_i),$$

where $\tilde{C}_\theta(x)$ is defined as

$$\min\left(\min_{m}{}_\tau \ \text{dist}(x, B_m) - \min_{m\colon y_m \neq y_{m^\star}}{}_\tau \ \text{dist}(x, B_m), 50\right). \tag{13}$$

Relaxing Indicator Functions Now observe that L_2 is still a function of indicator functions, due to the dist function in (13), which was derived in Lemma 4.1 to be $\text{dist}(x, B(a, b)) = \sum_{i=1}^{n} \mathbf{1}\left(x_i \notin [a_i, b_i]\right)$. Again, as the gradients of $\mathbf{1}[\cdot]$ are zero almost everywhere, we perform a conical approximation to $\mathbf{1}\left(x_i \notin [a_i, b_i]\right)$ which has non-zero gradients:

$$\text{conical}(x, a_i, b_i) \stackrel{\text{def}}{=} \max(a_i - x, 0) + \max(x - b_i, 0).$$

Finally, we replace the indicator $\mathbf{1}[f_\theta(x_i) = y_i]$ in L_2 by s_i, where $s_i = +1$ if $f(x_i) = y_i$, and $s_i = -1$ otherwise, to have the misclassified data-points contribute to the loss. These modifications lead to our final objective $L(\theta)$.

Improving Initialization We initialize θ by using a set of boxes defined from the data. This is done by first drawing a subset T of size M uniformly at random from the training data-points, and then initializing θ with axis-aligned boxes centered at these data-points, as $\{(B(x - 0.1, x + 0.1), y) \colon (x, y) \in T\}$, where $+$ denotes vector-scalar addition. Having described all the tricks used for optimizing Box-NN, we now proceed to performing an empirical evaluation.

5 Empirical Evaluation

In this section, we will briefly describe existing methods for probabilistic ℓ_0 certification, (Levine & Feizi, 2020b) and (Jia et al., 2022) as well as deterministic ℓ_0 certification (Hammoudeh & Lowd, 2023), and then empirically compare our (deterministic) ℓ_0 certified defense Box-NN to these approaches.

Levine & Feizi (2020b) and Jia et al. (2022) extend the technique of randomized smoothing (Cohen et al., 2019) to randomized ablation (RA), where given any classifier f (*e.g.*, a neural network), they produce a smoothed classifier g by zeroing out k pixels uniformly at random:

$$g_{\text{RA}}(x) = \arg\max_k \mathop{\mathbb{P}}_{v \sim \text{Unif}(S)} \left(f(x \odot v) = k\right), \tag{14}$$

where $S = \{v \in \{0, 1\}^n \colon \|v\|_0 = n - \rho\}$ is the discrete set of all binary vectors of length n having exactly ρ zeros, and $\odot$ denotes the Hadamard product. For this construction in (14), a counting argument leads to the

[3] $\nabla_c \min(c_1, c_2, \ldots, c_m) = (0, \ldots, 0, 1, 0, \ldots, 0) = e_{j^\star}$, where e_j is the jth standard basis vector, and $j^\star = \arg\min_j c_j$.

robustness certificate in Levine & Feizi (2020b), which we compare to in Fig. 2. A more complicated analysis based on the Neyman-Pearson lemma leads to a tighter certificate in Jia et al. (2022), which is also included in our comparison in Fig. 3 (left).Both these certificates are randomized, *i.e.*, they hold with a confidence $1-\alpha$, where α, ρ are hyper-parameters that trade-off benign accuracy to robustness, and can be chosen empirically. According to standard practice, we fix $\alpha = 0.05$ and produce plots for varying ρ. The interested reader can refer to (Levine & Feizi, 2020b; Jia et al., 2022) for a detailed description of these certification procedures.

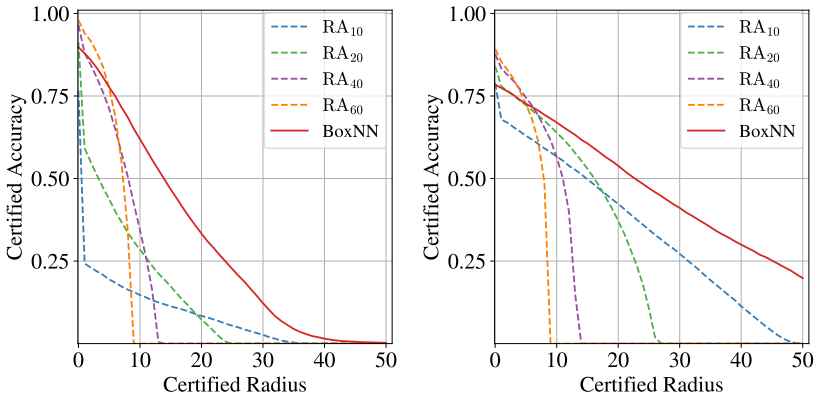

Figure 2: Comparison of Randomized Ablation (Levine & Feizi, 2020b) to our method Box-NN on the MNIST (left) and FashionMNIST (right) datasets. In each figure, the dotted lines correspond to different hyperparameter settings ρ. Details in text.

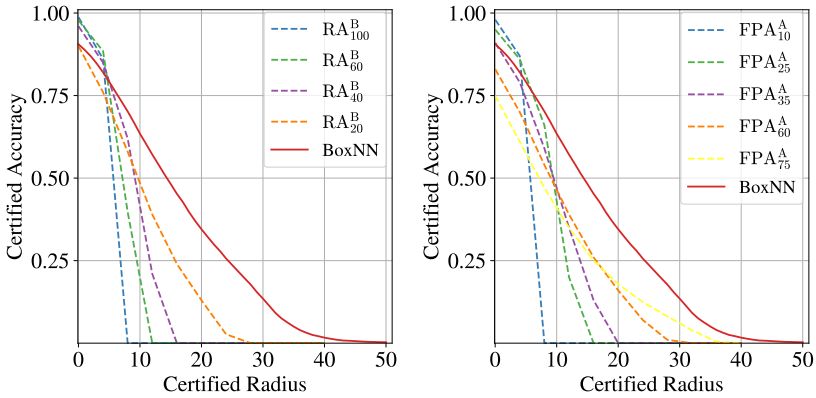

Figure 3: Comparison of Jia et al. (2022) (left) and Hammoudeh & Lowd (2023) (right) to our method Box-NN on the MNIST dataset. The dotted lines correspond to different settings for the hyperparameter ρ. Details are mentioned in text.

More recently, given any classifier f, Hammoudeh & Lowd (2023) produce a determinstic ℓ_0 certified classifier g by partitioning the set of pixels $\{1, 2, \ldots, n\}$ into disjoint partitions $\mathcal{S}$, and then producing the majority prediction of f over $\mathcal{S}$:

$$g_{\text{FPA}}(x) = \text{Majority}\{f(x_S)\}_{S \in \mathcal{S}}, \tag{15}$$

where $f(x_S)$ is defined as the prediction of f obtained after zeroing out the pixels in x not in S. Hammoudeh & Lowd (2023) then produce a certificate by counting the difference in the votes of the majority label to the runner-up label in (15). In Fig. 3 (right), we compare to the best performing strategy for constructing $\mathcal{S}$ in (Hammoudeh & Lowd, 2023) named "strided", where equally spaced pixels are selected for each partition,

Table 1: Comparison of the median certified radius $\bar{r}$ obtained by our Box-NN to the best hyperparameter settings for prior work.

Dataset	Method	$\bar{r}$
MNIST	Box-NN	**13**
	RA Levine & Feizi (2020b)	8
	RA^{B} Jia et al. (2022)	10
	FPA^{A} Hammoudeh & Lowd (2023)	9
	FPA^{B} Hammoudeh & Lowd (2023)	12
FMNIST	Box-NN	**22**
	RA Levine & Feizi (2020b)	16

i.e., $\mathcal{S} = \{p \colon p \equiv t - 1 \bmod \rho\}_{t=0}^{\rho-1}$. Here ρ is a hyper-parameter as earlier, and we vary ρ to produce the plots in Fig. 3[4]. Note that (Hammoudeh & Lowd, 2023) also obtain an improved certificate by using an aggregation more complicated than the majority vote, which we compare to in Fig. 4. The interested reader can refer to Appendix B and (Hammoudeh & Lowd, 2023) for more details.

Results Recall from Section 4.2 Eq. (10) that the certified accuracy of a classifier g against ϵ-bounded adversarial perturbations, $\text{CertAcc}(g, \epsilon)$, can be obtained given a point-wise certificate C for g. For each of the methods described so far, we plot CertAcc against ϵ using the corresponding robust classifier g and the certificate C over samples from the test set of the datasets mentioned.

A commonly used metric for comparing certified accuracy curves adopted in the literature (Levine & Feizi, 2020b; Jia et al., 2022; Hammoudeh & Lowd, 2023) is the median certified radius, which is the largest perturbation strength under which a classifier is certified to have atleast 50% robust accuracy. As can be seen in Table 1, our method Box-NN *outperforms all existing methods under all hyperparameter settings on this metric.*

The median certified radius captures a small slice of the full certified accuracy curve, which provides a complete picture. Observe that the dotted curves in Figs. 2 and 3 remain lower than our red curve except at small attack strengths. This shows that Box-NN is able to produce better certificates at most radii, and trades-off robustness at higher radii for benign accuracy at small radii. Without any dedicated hyper-parameter tuning, Box-NN dominates any single dotted curve for a large range of attack strengths, demonstrating that certified defenses closely utilizing properties of the data-distribution can outperform complicated ensembling-based defenses which ignore properties of the data.

6 Conclusion, Limitations and Future Work

In this work, we developed a theoretical to exploit properties of the data distribution for robustness against sparse adversarial attacks. We showed that data localization – the property that a data distribution p places most of its mass on very small volume sets in the input space – characterizes the existence of a ℓ_0-robust classifier for p. Following this theory, we developed a defense against sparse adversarial attacks, and derived a corresponding robustness certificate. We showed that this certificate empirically improves upon existing state-of-the-art in several broad regimes.

The primary limitation of our work is the difficulty in efficiently learning classifiers that have axis-aligned decision regions. While we are able to successfully employ several optimization tricks for datasets like MNIST and Fashion MNIST, the task becomes harder on more complicated datasets, *even though the geometry required for the underlying data-distribution remains the same due to our general theoretical results*. These optimization difficulties mostly stem from the strict requirement of axis-aligned boxes for our distance computation in Lemma 4.1. In the future, we hope to trade-off efficiency in the distance computation in favor of richer decision boundaries that can be learnt efficiently and generalize well.

[4]We use the results reported in Hammoudeh & Lowd (2023, Table 27) given that no public implementation of the method is available, to the best of our knowledge.

References

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Zheng Dai and David Gifford. Fundamental limits on the robustness of image classifiers. In *The Eleventh International Conference on Learning Representations*, 2022.

Elvis Dohmatob. Generalized no free lunch theorem for adversarial robustness. In *International Conference on Machine Learning*, pp. 1646–1654. PMLR, 2019.

Francisco Eiras, Motasem Alfarra, Philip Torr, M Pawan Kumar, Puneet K Dokania, Bernard Ghanem, and Adel Bibi. Ancer: Anisotropic certification via sample-wise volume maximization. *Transactions of Machine Learning Research*, 2022.

Marc Fischer, Maximilian Baader, and Martin Vechev. Certified defense to image transformations via randomized smoothing. *Advances in Neural Information Processing Systems*, 33:8404–8417, 2020.

Zayd Hammoudeh and Daniel Lowd. Feature partition aggregation: A fast certified defense against a union of sparse adversarial attacks. *arXiv preprint arXiv:2302.11628*, 2023.

Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. *Advances in Neural Information Processing Systems*, 33:10558–10570, 2020.

Jinyuan Jia, Binghui Wang, Xiaoyu Cao, Hongbin Liu, and Neil Zhenqiang Gong. Almost tight l0-norm certified robustness of top-k predictions against adversarial perturbations. In *International Conference on Learning Representations*, 2022.

Alexander Levine and Soheil Feizi. (de) randomized smoothing for certifiable defense against patch attacks. *Advances in Neural Information Processing Systems*, 33:6465–6475, 2020a.

Alexander Levine and Soheil Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020b.

Ambar Pal and Jeremias Sulam. Understanding noise-augmented training for randomized smoothing. *Transactions on Machine Learning Research*, 2023.

Ambar Pal and Rene Vidal. A game theoretic analysis of additive adversarial attacks and defenses. *Advances in Neural Information Processing Systems*, 2020.

Ambar Pal, Jeremias Sulam, and René Vidal. Adversarial examples might be avoidable: The role of data concentration in adversarial robustness. *Advances in Neural Information Processing Systems*, 2023.

Samuel Pfrommer, Brendon G Anderson, and Somayeh Sojoudi. Projected randomized smoothing for certified adversarial robustness. *Transactions on Machine Learning Research*, 2023.

Hadi Salman, Saachi Jain, Eric Wong, and Aleksander Madry. Certified patch robustness via smoothed vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15137–15147, 2022.

Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Michael Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81:73–205, 1995.

Michel Talagrand. A new look at independence. *The Annals of probability*, pp. 1–34, 1996.

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, 2018.

A Auxilliary Lemmas and Proofs

Lemma A.1 (Properties of expansion and contraction, extending Pal et al. (2023))**.** *For a distance d, set $A \subseteq [0,1]^n$, define $A^{+\epsilon} = \{x \in [0,1]^n \colon \mathrm{dist}_d(x, A) \leq \epsilon\}$, and $A^{-\epsilon} = \{x \in [0,1]^n \colon B_d(x,\epsilon) \subseteq A\}$. Then, for $N, O \subseteq [0,1]^n$, we have*

1. $(N \cap O)^{-\epsilon} = N^{-\epsilon} \cap O^{-\epsilon}$
2. $(N^c)^{-\epsilon} = (N^{+\epsilon})^c$, *where c denotes complement in* $[0,1]^n$
3. $(N \setminus O)^{-\epsilon} = N^{-\epsilon} \setminus O^{+\epsilon}$
4. $(N \cup O)^{+\epsilon} = N^{+\epsilon} \cup O^{+\epsilon}$
5. $(N^{+\epsilon_1})^{+\epsilon_2} \subseteq N^{+(\epsilon_1+\epsilon_2)}$

Proof. The first four assertions of this Lemma are standard results in mathematical morphology, dealing with the erosion and dilation of sets, and are reproduced here from Pal et al. (2023) for clarity.

1. Let $M = N \cap O$.

$$\begin{aligned} M^{-\epsilon} &= \{x \colon x \in M, B_d(x,\epsilon) \subseteq M\} \\ &= \{x \colon x \in N, x \in O, B_d(x,\epsilon) \subseteq N, B_d(x,\epsilon) \subseteq O\} = N^{-\epsilon} \cap O^{-\epsilon}. \end{aligned}$$

2. Let $M = N^c$.

$$\begin{aligned} M^{-\epsilon} &= \{x \colon x \in M, B_d(x,\epsilon) \subseteq M\} = \{x \colon x \notin N, B_d(x,\epsilon) \subseteq N^c\} \\ &= \{x \colon x \notin N, \forall x' \in B_d(x,\epsilon)\ x' \notin N\} \\ &= \{x \colon \forall x' \in B_d(x,\epsilon)\ x' \notin N\} \\ \implies (M^{-\epsilon})^c &= \{x \colon \exists x' \in B_d(x,\epsilon)\ x' \in N\} \\ &= N^{+\epsilon}. \end{aligned}$$

3. Let $M = N \setminus O$, we have $M^{-\epsilon} = (N \cap O^c)^{-\epsilon} = N^{-\epsilon} \cap (O^c)^{-\epsilon}$ by Property 1, and then $N^{-\epsilon} \cap (O^c)^{-\epsilon} = N^{-\epsilon} \cap (O^{+\epsilon})^c$ by Property 2.
4. Let $M = N \cup O$. We have $M^c = N^c \cap O^c$. Taking ϵ-contractions, and applying the first and second properties, we get $M^{+\epsilon} = N^{+\epsilon} \cup O^{+\epsilon}$.
5. For a set M, and any $\epsilon_1 \geq 0, \epsilon_2 \geq 0$, we have

$$\left(M^{+\epsilon_1}\right)^{+\epsilon_2} \subseteq M^{+(\epsilon_1+\epsilon_2)}.$$

The above property can be derived from the triangle inequality applied to d, as

$$\begin{aligned} \left(M^{+\epsilon_1}\right)^{+\epsilon_2} &= \{x \colon \exists x' \in M^{+\epsilon_1},\ d(x', x) \leq \epsilon_2\} \\ &= \{x \colon \exists x' \in \mathcal{X}, x'' \in M,\ d(x', x) \leq \epsilon_2, d(x'', x') \leq \epsilon_1\} \\ &\subseteq \{x \colon \exists x'' \in M,\ d(x'', x) \leq \epsilon_2 + \epsilon_1\} = M^{+(\epsilon_1+\epsilon_2)}. \end{aligned}$$

□

Lemma 4.1 (ℓ_0 distance to axis-aligned boxes)**.** *For an axis aligned box $B(a,b)$ specified as $B(a,b) = \{x \colon a \leq x \leq b\}$, where $a, b, x \in \mathbb{R}^n$, and all inequalities are element-wise, we have*

$$\mathrm{dist}(x, B(a,b)) = \sum_{i=1}^{n} \mathbf{1}\left(x_i \notin [a_i, b_i]\right),$$

which can be computed in $O(n)$ operations.

Proof. For any given x, recall the definition of dist to be $\mathrm{dist}(x, B(a,b)) = \min_{y \in B(a,b)} \|x - y\|_0$. For any $y \in B(a,b)$ we have,

$$\|x - y\|_0 = \sum_{i=1}^{n} \mathbf{1}(x_i \neq y_i) \geq \sum_{i=1}^{n} \mathbf{1}(x_i \notin [a_i, b_i])\mathbf{1}(y_i \in [a_i, b_i]) = \sum_{i=1}^{n} \mathbf{1}(x_i \notin [a_i, b_i]) \tag{16}$$

The above implies $\min_{y \in B(a,b)} \|x - y\|_0 \geq \sum_{i=1}^{n} \mathbf{1}(x_i \notin [a_i, b_i])$. Then, consider $y^\star \in B(a,b)$ defined as

$$y_i^\star = \begin{cases} a_i & \text{if } x_i \notin [a_i, b_i] \\ x_i & \text{otherwise} \end{cases}. \tag{17}$$

We have $\|y^\star - x\|_0 = \sum_{i=1}^{n} \mathbf{1}(x_i \notin [a_i, b_i])$, which attains the lower bound on $\mathrm{dist}(x, B(a,b))$. The result follows. □

Theorem 4.2 (Robustness Certificate for Box-NN)**.** *Given a set of boxes $\mathcal{B}$ and their associated labels $\{y_m\}_{m=1}^{M}$, define*

$$m^\star = \arg\min_m \mathrm{dist}(x, B_m), \quad d_1 = \mathrm{dist}(x, B_{m^\star}),$$

and

$$d_2 = \min_{m\colon y_m \neq y_{m^*}} \mathrm{dist}(x, B_m).$$

Then, with $\mathrm{margin}(x) \stackrel{\text{def}}{=} d_2 - d_1$*, we have* Box-NN$(x, \mathcal{B})$ = Box-NN$(x', \mathcal{B})$ *whenever* $\|x' - x\|_0 < \mathrm{margin}(x)/2$.

Proof. Let $x, x' \in \mathcal{X}$. Define $\mathcal{B}_1 = \{B_m \colon y_m = y_{m^\star}\}$, and $\mathcal{B}_2 = \{B_m \colon y_m \neq y_{m^\star}\}$. Further, define $\bar{d}_1, \bar{d}_2$ as

$$d_1(x') = \min_{B \in \mathcal{B}_1} \mathrm{dist}(x', B), \quad d_2(x') = \min_{B \in \mathcal{B}_2} \mathrm{dist}(x', B),$$

Our goal would be to demonstrate that as long as $\|x - x'\|_0 < \mathrm{margin}(x)/2$, we have $d_2(x') > d_1(x')$, implying that the prediction remains the same at x'. Consider any $B \in \mathcal{B}_2$, and apply the triangle inequality to get

$$\mathrm{dist}(x', B) + \|x - x'\|_0 \geq \mathrm{dist}(x, B), \tag{18}$$

where (18) can be seen as

$$\mathrm{dist}(x', B) + \|x - x'\|_0 = \min_{y \in B} \|y - x'\|_0 + \|x' - x\|_0 \geq \min_{y \in B} \|y - x\|_0 = \mathrm{dist}(x, B). \tag{19}$$

Further, taking a minimum on both sides of (18) over all $B \in \mathcal{B}_2$ leads to

$$d_2(x') + \|x - x'\|_0 \geq d_2 \tag{20}$$

Similarly, consider any $B \in \mathcal{B}_1$, and apply the triangle inequality to get

$$\mathrm{dist}(x, B) + \|x - x'\|_0 \geq \mathrm{dist}(x', B), . \tag{21}$$

Taking a minimum over both sides of (21) over all $B \in \mathcal{B}_1$ leads to

$$d_1 + \|x - x'\|_0 \geq d_1(x'). \tag{22}$$

Adding (20) and (22), we have

$$d_2(x') - d_1(x') + 2\|x - x'\|_0 \geq d_2 - d_1 \tag{23}$$

$$\implies d_2(x') - d_1(x') \geq \mathrm{margin}(x) - 2\|x - x'\|_0, \tag{24}$$

from where we can see that $d_2(x') - d_1(x') > 0$ whenever $\|x - x'\|_0 < \mathrm{margin}(x)/2$, as required. □

B Additional Empirical Comparison

We produce an additional comparison to ℓ_0 certificates in Hammoudeh & Lowd (2023). Since there is no publicly available code for this method, we compare our method against the numbers reported in Hammoudeh & Lowd (2023, Table 27). In Fig. 4, we compare against the method "FPA with run-off elections" reported in Hammoudeh & Lowd (2023). This method uses a more complicated aggregation scheme on top of Eq. (15) to obtain improved certificates. Nevertheless we observe that Box-NN improves upon the median certified robustness for all methods in all hyper parameter settings.

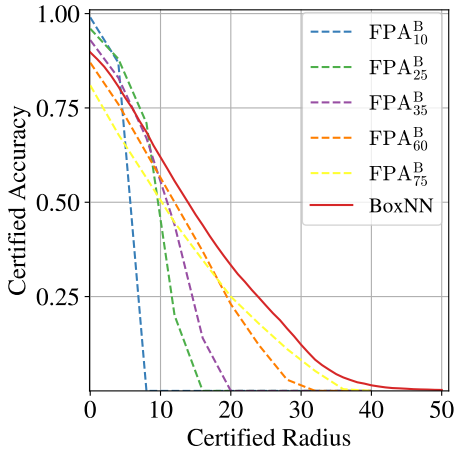

Figure 4: Comparison of a deterministic certificate Hammoudeh & Lowd (2023) (dotted lines) to our method Box-NN (red line) on the MNIST dataset. The dotted lines correspond to different settings for the hyperparameter ρ. Details are mentioned in main text.