Comparing Complexities of Decision Boundaries for Robust Training: A Universal Approach

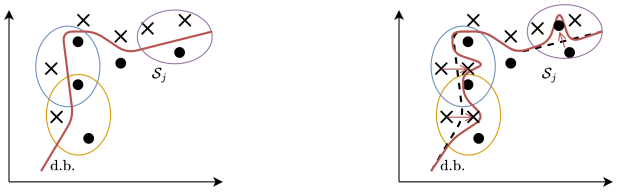
Daniel Kienitz, Ekaterina Komendantskaya, and Michael Lones

Heriot-Watt University, Edinburgh, UK
{dk50, e.komendantskaya, m.lones}@hw.ac.uk

Abstract. We investigate the geometric complexity of decision boundaries for robust training compared to standard training. By considering the local geometry of nearest neighbour sets, we study them in a model-agnostic way and theoretically derive a lower-bound $R^* \in \mathbb{R}$ on the perturbation magnitude $\delta \in \mathbb{R}$ for which robust training provably requires a geometrically more complex decision boundary than accurate training. We show that state-of-the-art robust models learn more complex decision boundaries than their non-robust counterparts, confirming previous hypotheses. Then, we compute R^* for common image benchmarks and find that it also empirically serves as an upper bound over which label noise is introduced. We demonstrate for deep neural network classifiers that perturbation magnitudes $\delta \geq R^*$ lead to reduced robustness and generalization performance. Therefore, R^* bounds the maximum feasible perturbation magnitude for norm-bounded robust training and data augmentation. Finally, we show that $R^* < 0.5R$ for common benchmarks, where R is a distribution's minimum nearest neighbour distance. Thus, we improve previous work on determining a distribution's maximum robust radius.

1 Introduction

The decision boundary learned by a classifier is a crucial property to study [1–3]. Its geometric complexity, i.e. its number of linear segments, is an indication of the train distribution's complexity and the difficulty of learning [4–6], its margin to the train samples defines its robustness [1, 2, 7] and studying its general position in input space is used for explaining model predictions [8]. Deep neural network classifiers have over the past years reached or even surpassed human-level performance in computer vision tasks [9, 10]. However, despite recent progress [11] they still remain vulnerable to a large variety of distribution shifts [12–16]. In light of this brittleness [12, 13, 17] and the observation that robust training methods cause non-robust accuracy to deteriorate [18, 17, 19–22], several recent works have hypothesized that robust training might require different, and possibly geometrically more complex, decision boundaries than standard training [23–26, 17, 1]. If this hypothesis is true, the need for greater capacity [26] and increased sample complexity of robust training (see Section 2 for overview) could partially be explained. As the decision boundary learned by



(a) Original. (b) Perturbed.

Fig. 1: Illustration of the main idea in $d = 2$ dimensions. (a) The input distribution is separated into sets $\mathcal{S}_j$ of linearly separable nearest neighbours. (b) If $\mathcal{S}_j$ changes and is no longer linearly separable, the complexity of the decision boundary (d.b.) increases.

a deep classifier is build on top of a largely opaque feature representation [23, 27], is high-dimensional, highly non-linear and could theoretically consist of multiple disconnected decision regions [28] its study is a challenging problem.

In this paper we take a model-agnostic approach to studying decision boundaries. We assume the existence of an *accurate decision boundary*, obtained by minimizing the train loss, which perfectly separates the input data's classes and give a comparative study of the *robust decision boundary* that would be required if the data was altered by worst-case perturbations of its samples. To achieve this, we divide the input distribution into linearly separable sets $\mathcal{S}$ of nearest neighbours and investigate the perturbation magnitudes required to make them non-linearly separable. This approach allows us to make the following contributions:

- On the theoretical side, we derive a lower bound $R^* \in \mathbb{R}$ in l_2-norm on the perturbation magnitude $\delta \in \mathbb{R}$ in l_2-norm in input space for which the geometric complexity of a robust decision boundary provably increases compared to an accurate decision boundary. (See Section 3).
- Since R^* is efficiently computable, we show for common image benchmarks that state-of-the-art robust deep classifiers indeed learn geometrically more complex decision boundaries than their accurate counter parts and that they are better calibrated in low-density regions. (See Section 4).
- When computing R^* for common image benchmarks, we find that perturbation magnitudes $\delta \geq R^*$ introduce label noise and demonstrate that this leads to decreased robustness and generalization performance. (See Section 5).

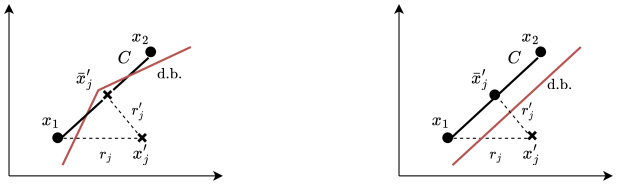
As the geometric complexity of decision boundaries is a crucial factor for the sample complexity of a learning problem [4–6], showing under which perturbation magnitudes decision boundaries increase in complexity is important for the selection of hypothesis classes. Further, as label noise is known to be one of the reasons for the lack of robustness [27], bounding the maximum perturbation magnitude for which norm-bounded robust training and data augmentation is possible for these benchmarks, is crucial in practical applications.

Finally, we show that R^* is a more accurate approximation of a distribution's robust radius. Previous work [29] utilized the minimum nearest neighbour distance between any two samples in the dataset, $R \in \mathbb{R}$. We show that $R^* < 0.5R$ for common image benchmarks.

2 Related Work

Learned representations of standard training It has been shown that part of the adversarial vulnerability of deep classifiers stems from common image benchmarks containing highly predictive yet brittle features [30, 31]. These features are usually not aligned with those used by humans for classification [32], so small-norm and non-semantic changes to inputs are often sufficient to change the classification decisions of otherwise well-performing classifiers [12, 13, 33]. Since neural networks were shown to rely on simple features for classification [34, 27, 35–37], even in the presence of complex ones with better predictive power [23], models were found to learn feature representations on top of superficial statistical regularities [38, 39] like texture [40–42] and non-semantic pixel subsets [43, 44]. As decision boundaries are functions of the learned representation, this simplicity bias and the presence of non-semantic but highly predictive features leads to boundaries that are accurate but not robust. These observations led several authors to suggested that robust training might require more complex decision boundaries than accurate training [24–26, 17, 1]. In this work we confirm this hypothesis and further show that in the low-density region where the decision boundary is supposed to lie [45], state-of-the-art robust classifiers are largely better calibrated than non-robust ones.

Sample complexity Several studies argue that adversarial training has a larger sample complexity than standard training. Bounds on the sample complexity where the data distribution is a mixture of Gaussians were first provided by Schmidt et al. [24] who showed that the increased sample complexity of adversarial training holds regardless of the training algorithm and the model family. Later, Bhagoji et al. [46] studied sample complexity with an approach from optimal transport. Dobriban et al. [47] extended prior analyses to mixtures of three Gaussians in 2- and ∞-norm and Dan et al. [48] derived general results for the case of two-mixture Gaussians for all norms. More recently, Bhattacharjee et al. [49] studied the sample complexity of robust classification for linearly separable datasets. They showed that in contrast to accurate classification, the sample complexity of robust classification has a linear dependence on the dimension d. Yin et al. [25] further showed a dependence of the sample complexity on d for neural networks. Distribution-agnostic bounds for robust classification have been provided by several authors [50–53]. As the sample complexity is also influenced by the geometric complexity of the decision boundary [4–6], we provide another reason for its increase for robust training by showing that robust models learn more complex decision boundaries compared to non-robust ones.



(a) $f(x'_j) = f(\bar{x}'_j)$. (b) $f(x'_j) \neq f(\bar{x}'_j)$.

Fig. 2: Illustration of an $\mathcal{S}_j$ in $d = 2$ dimensions. (a) If classifier f does not assign a class change between x'_j and $\bar{x}'_j$, the decision boundary's (d.b.) complexity increases. (b) If f assigns a class change to $\bar{x}'_j$, $\mathcal{S}_j$ is still linearly separable, however, if $r'_j < 0.5r_j$, robust training for $\delta = r'_j$ introduces label noise.

Robust training Recently, several methods have been proposed to mitigate the reliance on superficial regularities by removing texture clues [54, 40], improved data augmentation [55, 56], pre-training [57] and utilization of unlabelled data in the training process [58, 59, 56]. Further, several robust training methods like adversarial training [13, 17], regularization [60, 61] and saliency methods [62–64] have been proposed. Nevertheless, all previously mentioned methods found reduced generalization performance with increasing robustness. In this work we hypothesize and empirically confirm that robust training for large magnitudes reduces generalization performance because it introduces label noise which biases the model towards learning non-generalizing textural features. We provide a lower bound R^* over which this provably occurs for common image benchmarks.

3 Derivation of R^*

In this section we describe how R^* is derived for an arbitrary distribution $X \in \mathbb{R}^{l \times d}$, where l is the number of samples and d is the ambient dimension[1]. Each sample $x'_j \in X$ is associated with a unique label $y'_j \in Y$, where c is the number of distinct classes in Y.

3.1 Nearest Neighbour Sets

Since determining the geometric complexity of a distribution's decision boundary is still an open problem, we study under which perturbation magnitudes sets of linearly separable nearest neighbours $\mathcal{S}$ become non-linearly separable. This approach allows us to treat the geometric complexity of the decision boundary for X as the unknown base case and only investigate the increase in complexity under the worst-case perturbation. Without loss of generality we describe the method for binary classification, so $c = 2$ classes. Later we describe how it can be extended to the multi-class case where $c > 2$.

[1] We assume that samples do not lie on a flat manifold, so are not perfectly collinear.

For every sample $x'_j \in X$, $j = 1, ..., l$, we consider its d nearest neighbours (according to l_2-norm) with a different class label. This results in a separation of the input distribution $X \in \mathbb{R}^{l \times d}$ into l subsets $\mathcal{S}_j = \{x_1, ..., x_d, x'_j\}$. The samples $x_1, ..., x_d$ are the ordered d-nearest neighbours with the other class label where x_1 is the closest and x_d the farthest. Since the cardinality of each $\mathcal{S}_j$ is equal to the VC-dimension of a linear classifier, so $|\mathcal{S}_j| = d + 1$, a single hyperplane is sufficient to separate x'_j from $\{x_i\}_{i=1}^d$ with perfect accuracy. With this approach we can investigate what perturbation magnitude is required so that the set $\mathcal{S}_j = \{x_1, ..., x_d, x'_j\}$ is not linearly separable any more. The linear separability property of $\mathcal{S}_j$ is violated if x'_j is projected onto the convex hull $\mathcal{C}(\{x_i\}_{i=1}^d)$ of its d nearest neighbours. We define the projection of x'_j onto $\mathcal{C}(\{x_i\}_{i=1}^d)$ as

$$\begin{aligned} \bar{x}'_j &\coloneqq \text{argmin}_{\hat{x} \in \mathcal{C}(\{x_i\}_{i=1}^d)} ||x'_j - \hat{x}||_2 \text{ s.t.} \\ \hat{x} &= \sum_{i=1}^{d} w_i x_i,\ 0 \leq w_i \leq 1,\ \sum_{i=1}^{d} w_i = 1 \end{aligned} \tag{1}$$

Thus, replacing x'_j with $\bar{x}'_j$ in $\mathcal{S}_j$ removes the linear separability property because all samples are collinear. In Appendix A we describe how the optimization problem in Equation 1 can be solved exactly and deterministically and show that choosing $|\mathcal{S}_j| \neq d + 1$ leads to a vacuous bound.

3.2 Properties of Nearest Neighbour Sets

We define the distance between a sample x'_j and its projection onto the convex hull of its nearest neighbours $\bar{x}'_j$ of another class as

$$r'_j \coloneqq ||x'_j - \bar{x}'_j||_2 \tag{2}$$

Further,

$$r_j \coloneqq ||x'_j - x_1||_2 \tag{3}$$

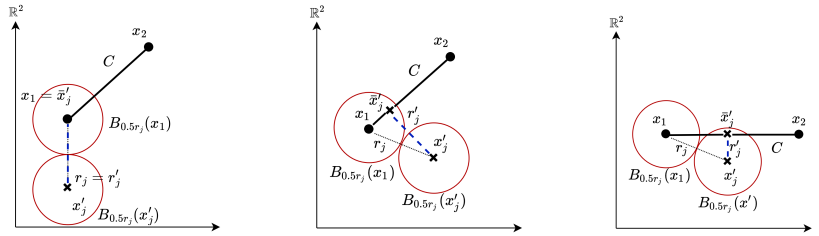
defines the distance between a sample x'_j and its nearest neighbour x_1 of a different class. The value $0.5r_j$ defines the same quantity as the robust radius defined by Yang et al. [29]. We illustrate these quantities in $d = 2$ dimensions in Figure 2.

3.3 Decision Boundary of Nearest Neighbour Sets

In Figure 3 we illustrate the main intuition behind our approach. Figuratively speaking, we require a more complex decision boundary for $\mathcal{S}_j$, so two connected hyperplanes instead of one[2], if the $(0.5r_j)$-ball $B_{0.5r_j}(x'_j) = \{x : ||x'_j - x||_2 \leq 0.5r_j\}$ of sample x'_j intersects with $\mathcal{C}(\{x_i\}_{i=1}^d)$. This is the case if $r'_j < 0.5r_j$. We define the threshold

$$r_j^{\text{crit}} \coloneqq \frac{r'_j}{0.5r_j} \tag{4}$$

[2] Note that enclosing just the point $\bar{x}'_j$ requires d hyperplanes arranged as a simplex.



(a) $r_j^{\text{crit}} = \frac{r'_j}{0.5r_j} = 2$ (b) $r_j^{\text{crit}} = \frac{r'_j}{0.5r_j} > 1$ (c) $r_j^{\text{crit}} = \frac{r'_j}{0.5r_j} < 1$

Fig. 3: Illustration of nearest neighbour sets $\mathcal{S}_j$ for $d = 2$ dimensions. (a) A more complex decision boundary is not required as $||x'_j - \bar{x}'_j||_2 = ||x'_j - x_1||_2 = r'_j = r_j$. (b) A more complex decision boundary is not required as $||x'_j - \bar{x}'_j||_2 = r'_j > 0.5r_j$. (c) A more complex decision boundary is required as $||x'_j - \bar{x}'_j||_2 = r'_j < 0.5r_j$.

for a single sample x'_j and its associated $\mathcal{S}_j$. If $r'_j < 0.5r_j$ then $r_j^{\text{crit}} < 1$ and robust accuracy with perturbation magnitudes of $\delta \geq r'_j$ *provably* requires a geometrically more complex decision boundary for $\mathcal{S}_j$. Conversely, $\mathcal{S}_j$ is still linearly separable for $r_j^{\text{crit}} > 1$ for any perturbation magnitude $\delta < 0.5r'_j$.

It is important to note that while the introduction of a perturbation $\delta < 0.5r'_j$ does not result in a locally more complex decision boundary for $\mathcal{S}_j$, it might result in a globally more complex decision boundary for the entire distribution X. Therefore, r'_j is the largest lower bound. Finding the smallest lower bound that holds globally requires knowing the optimal decision boundary which is generally unknown. As illustrated in Figure 1b, we assume that a locally more complex decision boundary for $\mathcal{S}_j$ results in a globally more complex decision boundary for X, so $\bar{x}'_j$ are not merely memorized. This assumption is reasonable as there is strong evidence that deep networks build connected decision regions encompassing all samples of a single class [65, 3].

Extension to multiple classes The method described above for the binary scenario can easily be extended to multi-class classification. Instead of determining the set of nearest neighbours $\mathcal{S}_j$ once for the single other class, the computation is repeated $(c-1)$-times for all other classes. The rationale from above holds, as we simply restrict the $B_{0.5r_j}(x'_j)$-ball to not intersect with any convex hull of nearest neighbours of any other class. So, the method scales linearly with the number of classes c in X. Then,

$$r'_j \coloneqq \min_{i \in Y \setminus y'_j} (\{r'_j(y_i)\}) \tag{5}$$

and

$$r_j \coloneqq \min_{i \in Y \setminus y'_j} (\{r_j(y_i)\}) \tag{6}$$

where $Y \setminus y'_j$ denotes the set of unique class labels without label y'_j of x'_j and $r'_j(y_i)$ and $r_j(y_i)$ denote the equivalents of r'_j and r_j defined in Equations 2 and

3, respectively, computed for class y_i. We always report the results for all classes in a particular dataset, unless stated otherwise.

Extension to the entire dataset The quantities in Equations 2, 3, and 4 are defined for a single $\mathcal{S}_j$. We define the robust radius of the entire distribution as

$$R \coloneqq \min_{j \in 1,...,l}(\{r_j\}) \tag{7}$$

which is equivalent to the definition of R by Yang et al. [29] and describes the minimum nearest neighbours distance between any two samples of different classes. Intuitively, $0.5R$ describes the maximum perturbation magnitude such that the $B_{0.5R}(\cdot)$-balls of any two samples do not intersect. We define

$$R_j^{\text{crit}} \coloneqq \frac{r'_j}{0.5 \cdot \min_{i \in 1,...,l}(\{r_i\})} = \frac{r'_j}{0.5R} \tag{8}$$

The interpretation of R_j^{crit} is equivalent to the one of r_j^{crit} in Equation 4 with the exception that we consider the global robust radius R instead of the local robust radius r_j. For $R_j^{\text{crit}} < 1$, the distance to the convex hull of a sample is smaller than the robust radius and therefore increases the complexity of the decision boundary. We further define

$$R^* \coloneqq \min_{j \in 1,...,l}(\{r'_j\}) \tag{9}$$

which describes the perturbation magnitude over which we provably require a geometrically more complex decision boundary for the given distribution. Finally,

$$R^{\text{crit}} \coloneqq \frac{\min_{j \in 1,...,l}(\{r'_j\})}{0.5 \cdot \min_{j \in 1,...,l}(\{r_j\})} = \frac{R^*}{0.5R} \tag{10}$$

describes whether R is an over-approximation of a distribution's robust radius.

Definition of critical points We refer to those points $\bar{x}'_j$ for which locally $r'_j < 0.5r_j$, so $r_j^{\text{crit}} < 1$, as *critical* as they require a locally more complex decision boundary under norm-bounded robustness scenarios and cause r_j to be an over-estimation of the actual robust radius,

$$\{\bar{x}\}_{\text{local}}^{\text{crit}} \coloneqq \{\bar{x}'_j : r_j^{\text{crit}} < 1,\ j = 1, ..., l\} \tag{11}$$

Conversely, we define those points for which $R_j^{\text{crit}} < 1$ as

$$\{\bar{x}\}_{\text{global}}^{\text{crit}} \coloneqq \{\bar{x}'_j : R_j^{\text{crit}} < 1,\ j = 1, ..., l\} \tag{12}$$

It follows that $|\{\bar{x}\}_{\text{global}}^{\text{crit}}| \leq |\{\bar{x}\}_{\text{local}}^{\text{crit}}|$. Note that in the multi-class case, $c > 2$, a single x'_j can have multiple associated $\bar{x}'_j$ that are elements of $\{\bar{x}\}_{\text{local}}^{\text{crit}}$ or $\{\bar{x}\}_{\text{global}}^{\text{crit}}$, possibly one for every other class in the dataset. Thus, $0 \leq |\{\bar{x}\}_{\text{global}}^{\text{crit}}| \leq l(c-1)$ and $0 \leq |\{\bar{x}\}_{\text{local}}^{\text{crit}}| \leq l(c-1)$, where l is the number of samples in X and c is the number of unique class labels.

3.4 Class Membership of $\bar{x}'$

Throughout this section we assumed that there is no change of the ground truth class between x'_j and its associated $\bar{x}'_j$. If Euclidean distance is a valid proxy for semantic similarity, then this assumption is valid. Furthermore, as only those $\bar{x}'_j$ for which its corresponding $r'_j < 0.5r_j$ (or $r'_j < 0.5R$) are of interest, the assumption of no class change is valid for such distributions. However, since it is well known that l_p-norms are not a suitable proxy for semantic similarity for real-world datasets, the class membership of $\bar{x}'_j$ cannot be inferred from a simple distance metric in input space. Thus, distinguishing between the three following possible scenarios is necessary.

No class change (NCC) If no change of the ground truth class label between x'_j and its corresponding $\bar{x}'_j$ occurs, then robust training for $\delta \geq r'_j$ requires a more complex decision boundary for $\mathcal{S}_j$ (see Figure 2a).

Class change (CC) If the ground truth class changes between x'_j and $\bar{x}'_j$, the decision boundary for $\mathcal{S}_j$ does not increase in complexity (see Figure 2b). However, this implies that R is not the actual robust radius of that distribution as there is at least one r_j for which $r'_j < 0.5r_j$. In this case, the robust radius is over-approximated by R and R^* is the actual robust radius. Crucially, this implies that robust training for magnitudes $\delta \geq R^*$ introduces label noise.

Ambiguous class If $\bar{x}'_j$ cannot be assigned a ground truth class membership, it lies within the low-density region between classes. In this case r_j is again an over-approximation of the actual robust radius and robust training for $\delta \geq R^*$ introduces label noise as well.

In summary, the interpretation of R^* depends on the class membership of $\bar{x}_j$. In all cases it is a model-agnostic lower-bound on the perturbation magnitude for which a geometrically more complex decision boundary is required. Therefore, it can guide the choice of hypothesis classes required for robust training. Further, it can also upper-bound the maximum feasible perturbation magnitude for a given dataset over which label noise is introduced. As label noise is known to hurt robustness [27], R^* can guide the usage of norm-bounded robust training and data augmentation for neural networks.

4 Computation of R^* for Image Benchmarks

In this section we compute R^* for real-world image benchmarks. We show that for those datasets it indeed upper-bounds the maximum feasible perturbation magnitude and that state-of-the-art robust models exhibit geometrically more complex decision boundaries. Finally, we also show that $R^* < 0.5R$, so it improves prior work on bounding the robust radius of a distribution [29].

As norm-bounded perturbations are usually given either in the l_2-norm or the l_∞-norm, we extend R^* to the l_∞-norm as well. Finally, we also show that R^* is independent of the ambient dimension d.

Table 1: Results in l_2-norm for real image benchmarks.

	R	$0.5R$	R^*	R^{crit}	$\lvert\{\bar{x}\}^{\text{crit}}_{\text{local}}\rvert$	$\frac{\lvert\{\bar{x}\}^{\text{crit}}_{\text{local}}\rvert}{l(c-1)}$	$\lvert\{\bar{x}\}^{\text{crit}}_{\text{global}}\rvert$	$\frac{\lvert\{\bar{x}\}^{\text{crit}}_{\text{global}}\rvert}{l(c-1)}$
SVHN	1.577	0.788	0.255	0.323	132061	0.200	2501	0.004
CIFAR-10	2.751	1.375	0.578	0.421	26608	0.059	132	0.000
FASHION	1.599	0.799	0.906	1.133	811	0.002	0	0.000
MNIST	2.398	1.199	1.654	1.379	0	0.000	0	0.000

Fig. 4: Example image-pairs of $\{\bar{x}\}^{\text{crit}}_{\text{global}}$ (right) their associated x'_j (left) for CIFAR-10. A single x'_j can be associated with multiple $\bar{x}'_j \in \{\bar{x}\}^{\text{crit}}_{\text{global}}$, possibly one for all other classes. Additional example images can be found in Appendix I.

4.1 R Overestimates the Robust Radius for Real Image Benchmarks

We compute the introduced quantities for the MNIST [68], FASHION [69], SVHN [70] and CIFAR-10 [71] datasets. For all datasets we use exact nearest neighbour search over the entire original train set. SVHN contains two mislabelled samples which we remove from the dataset (see Appendix E).

In Table 1, we display all the derived quantities from Section 3 for all aforementioned datasets. They display intuitive results. They confirm, for instance, the common knowledge that MNIST and FASHION are well-separated. As $R^* > 0.5R$, the robust radius is accurately described by $0.5R$. However, as R^* defines a lower bound, no definitive statement can be made about increases in the geometric complexity of the decision boundaries for robust training.

For the more sophisticated benchmarks SVHN and CIFAR-10 we observe that the nearest neighbour distance R is an overestimation of the actual robust radius, as $R^* < 0.5R$ and thus $R^{\text{crit}} < 1$. As a result, for both datasets $\{\bar{x}\}^{\text{crit}}_{\text{global}}$ are non-empty and it follows that they require a locally more complex decision boundary for perturbation magnitudes δ with $0.5R \geq \delta \geq R^*$.

$\bar{x}'_j$ are low-density samples In Section 3 we showed that the exact interpretation of R^* relies on the ground truth class label of the projections $\bar{x}'_j$. The question of class membership cannot be answered by l_p-norm distance metrics as for real-world datasets they are usually a bad proxy for semantic similarity. Thus, in Figure 4 we display several images from $\{\bar{x}\}^{\text{crit}}_{\text{global}}$ with their associated x'_j from the CIFAR-10 dataset. We find that the majority of $\bar{x}'_j$ are strongly blurred versions of their corresponding x'_j and do not contain a clearly recognizable object. Therefore, those samples are part of the low-density region between classes.

Table 2: Predictions and confidences of model f for $\{\bar{x}\}^{\text{crit}}_{\text{global}}$ for CIFAR-10. Confidence values are reported as: mean $\pm$ standard deviation. NCC denotes no predicted class change by f and CC denotes a predicted class change between x'_j and $\bar{x}'_j \in \{\bar{x}\}^{\text{crit}}_{\text{global}}$. The complete table can be found in Appendix I.

Model f		$f(x'_j) = f(\bar{x}'_j)$ (NCC)		$f(x'_j) \neq f(\bar{x}'_j)$ (CC)	
		Fraction	Confidence	Fraction	Confidence
Andriushchenko et al. [66]	Non-robust	0.62	0.887 ± 0.154	0.38	0.708 ± 0.192
	Robust	0.79	0.535 ± 0.164	0.21	0.373 ± 0.098
Ding et al. [7]	Non-robust	0.52	0.913 ± 0.171	0.48	0.826 ± 0.174
	Robust	0.93	0.979 ± 0.057	0.07	0.791 ± 0.113
Rebuffi et al. [56]	Non-robust	0.50	0.844 ± 0.177	0.50	0.650 ± 0.171
	Robust	0.94	0.643 ± 0.202	0.06	0.389 ± 0.083
Rice et al. [67]	Non-robust	0.54	0.863 ± 0.168	0.46	0.677 ± 0.204
	Robust	0.92	0.635 ± 0.200	0.08	0.401 ± 0.072

4.2 Robust Models Learn more Complex Decision Boundaries

In addition to the visual investigation of $(x'_j, \bar{x}'_j)$, we gather the predictions and confidences of thirteen state-of-the-art robust models on $\{\bar{x}\}^{\text{crit}}_{\text{global}}$ from CIFAR-10. These models are obtained from Croce et al. [72] and referred to as *robust* models. In addition, we re-initialize these architectures and re-train only with the Adam optimizer [73] on the original train set to remove their robust representation. Thus, the re-trained models are their *non-robust* counterparts. In Table 2 we display, due to space constraints, four models. For all thirteen models we observe two major differences between the robust and non-robust ones. Firstly, the non-robust models assign high confidences to $\{\bar{x}\}^{\text{crit}}_{\text{global}}$. As the visual inspection shows that $\{\bar{x}\}^{\text{crit}}_{\text{global}}$ are part of the low-density region between classes, high confidence scores indicate a poorly calibrated classifier. In contrast, the robust models usually assign significantly lower confidences to these low-density samples, a result that would be expected from a well-performing classifier. Secondly, we find that robust and non-robust models differ in their predictions of whether a class change has occurred between x'_j and its corresponding $\bar{x}'_j$. Whereas the robust models predict in most of the cases that no class change occurs, the non-robust models predict class changes in half of the cases. As the addition of $\{\bar{x}\}^{\text{crit}}_{\text{global}}$ to the train set increases the geometric complexity of the decision boundary, robust models learn more complex decision boundaries. Thus, we experimentally confirm the previously made hypothesis [24–26, 17, 1]. These results also partially explain why robust training has a greater sample complexity than standard training, since the geometric complexity of decision boundaries is known to increase the sample complexity [4–6].

4.3 From l_2- to l_∞-norm

Perturbation magnitudes for robust training and data augmentation are usually given in l_2- or l_∞-norm. In the previous section we computed R^* in l_2-norm so

Table 3: Number of pixels $\lceil (p-\tilde{p}) \rceil$ (see Equation 13) that need to be perturbed by ϵ in l_∞-norm to introduce perturbations $\delta > R^*$ in l_2-norm.

	$0.5R$	R^*	p	$\lceil (p-\tilde{p}) \rceil$				
				$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	$\epsilon = \frac{8}{255}$	$\epsilon = \frac{16}{255}$	$\epsilon = \frac{32}{255}$
SVHN	1.003	0.525	1024	4478	1120	280	70	18
CIFAR-10	1.375	0.578	1024	5439	1360	340	85	22
FASHION	0.799	0.906	784	13340	3335	834	209	53
MNIST	1.199	1.654	784	44461	11116	2779	695	174

here we expand the analysis to the l_∞-norm. Since the l_∞-norm is the maximum absolute change ϵ between any two vector dimensions, we compute how many dimensions in common image benchmarks need to be changed to surpass the specific R^*-value in l_2-norm. It is common practice in the robustness literature to apply the l_∞-norm on the pixel level, so to ignore the colour channel. Denoting the number of pixels in a dataset as p, with $0 \leq \tilde{p} \leq p$, it is easy to see that

$$\begin{aligned} ||x - \tilde{x}||_2 &= \sqrt{\sum_{i=1}^{p} (x_i - \tilde{x}_i)^2} = \sqrt{\sum_{i=1}^{\tilde{p}-1} (x_i - \tilde{x}_i)^2 + \sum_{j=\tilde{p}}^{p} (x_j - \tilde{x}_j)^2} > R^* \\ &= \sqrt{\sum_{j=\tilde{p}}^{p} \epsilon^2} > R^* \Leftrightarrow (p - \tilde{p}) > \left(\frac{R^*}{\epsilon}\right)^2 \end{aligned} \tag{13}$$

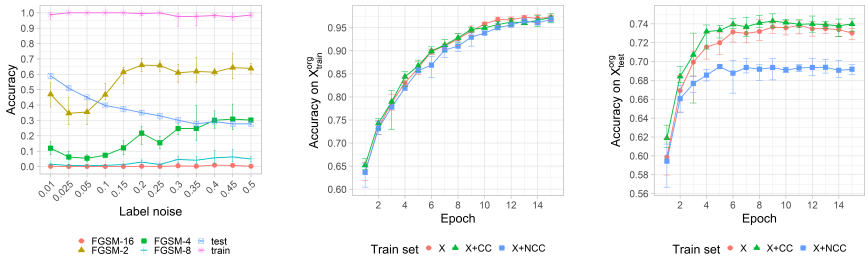
where the first sum is equal to zero because those pixels are not altered and the pixels in the second sum are all changed by ϵ due to the l_∞-norm. Thus, $(p - \tilde{p})$ is the number of pixels that need to be changed by ϵ such that the resulting perturbation magnitude in l_2-norm surpasses R^*. We round $(p - \tilde{p})$ to the nearest integer. In Table 3 we display the minimum number of pixels $\lceil (p-\tilde{p}) \rceil$ that need to be changed to surpass R^* in l_2-norm when perturbations ϵ are applied in l_∞-norm. For CIFAR-10, for example, we observe that a l_∞ perturbation magnitude of $\epsilon = 4/255$ requires $1,360$ pixels to be altered. As this is more than the original number of $1,024$ pixels, R^* is not surpassed in l_2-norm. In general, we observe that for the common perturbation magnitude $\epsilon = 8/255$ only a small fraction of pixels need to be altered in both SVHN and CIFAR-10 to surpass the threshold R^* in l_2-norm. In Section 5 we show that including samples with perturbation magnitude $\delta \geq R^*$ leads to reduced generalization performance.

4.4 Results are Independent of the Ambient Dimension *d*

The point $\bar{x}'_j$ minimizes the Euclidean distance between x'_j and the convex hull $\mathcal{C}(\{x_i\}_{i=1}^d)$. Since the convex hull is defined by the d nearest neighbours with

Table 4: Influence of the ambient dimension d for CIFAR-10.

	d	R	$0.5R$	R^*	R^{crit}	$\lvert\{\bar{x}\}^{\text{crit}}_{\text{local}}\rvert$	$\frac{\lvert\{\bar{x}\}^{\text{crit}}_{\text{local}}\rvert}{l(c-1)}$	$\lvert\{\bar{x}\}^{\text{crit}}_{\text{global}}\rvert$	$\frac{\lvert\{\bar{x}\}^{\text{crit}}_{\text{global}}\rvert}{l(c-1)}$
34x34	3468	2.843	1.422	0.497	0.350	38827	0.086	283	0.001
32x32	3072	2.751	1.375	0.578	0.421	26608	0.059	132	0.000
30x30	2700	2.514	1.257	0.486	0.386	36167	0.080	236	0.001
28x28	2352	2.324	1.162	0.467	0.402	34784	0.077	214	0.000
26x26	2028	2.157	1.079	0.444	0.412	33066	0.073	200	0.000



(a) (b) (c)

Fig. 5: Results for CIFAR-10. Error-bars denote minimum and maximum over five runs. (a) Mean accuracy on the train and test set and against FGSM-i/255, $i \in \{2, 5, 8\}$ attacks for different levels of label noise introduced by $\{\bar{x}\}^{\text{crit}}_{\text{global}}$. (b) Mean accuracy on $X^{\text{org}}_{\text{train}}$ during training with $\{\bar{x}\}^{\text{crit}}_{\text{local}}$. (c) Mean accuracy on $X^{\text{org}}_{\text{test}}$ during training with $\{\bar{x}\}^{\text{crit}}_{\text{local}}$.

another class label of x'_j, all quantities that are deducted from $\bar{x}'_j$ are functions of the ambient dimension d. Therefore, we investigate whether changes of the ambient dimension change the previously computed quantities.

We report the results for CIFAR-10 in Table 4. For image distributions an increase in their ambient dimension d, so their resolution, results in higher correlations between pixels and larger Euclidean distances between images. So, simultaneously higher values of R and R^* are expected. Further, there is no clear relationship between $\lvert\{\bar{x}\}^{\text{crit}}_{\text{local}}\rvert$ and $\lvert\{\bar{x}\}^{\text{crit}}_{\text{global}}\rvert$ with respect to d. Thus, the derived quantities are not artefacts of high dimensional spaces but dataset specific properties.

5 Training with $\delta \geq R^*$ Deteriorates Performance

In Section 3 we derived R^* theoretically for arbitrary datasets. We discussed that the implications of robust training for perturbation magnitudes $\delta \geq R^*$ depend on the class membership of those samples for which $r'_j < 0.5r_j$. In Section 4 we showed that for sophisticated real-world benchmarks $\{\bar{x}\}^{\text{crit}}_{\text{global}}$ lie within the low-density region between classes. Including these samples that do not display

Table 5: Accuracy against noise- (top) and blur-perturbations (bottom) [14] for CIFAR-10 and $\{\bar{x}\}^{\text{crit}}_{\text{local}}$. Networks trained on X+CC and X+NCC exhibit better robustness against small-norm noise-perturbations but worse robustness against large-norm blur perturbations. Example images can be found in Appendix H.

	Gaussian	Shot	Impulse	Speckle
X	0.521	0.501	0.537	0.557
X+CC	0.526	0.504	0.546	0.562
X+NCC	**0.600**	**0.581**	**0.584**	**0.621**

	Zoom	Defocus	Gaussian	Glass	Fog	Brightness	Contrast
X	**0.638**	**0.691**	0.324	0.703	**0.352**	0.694	0.328
X+CC	0.625	0.689	0.273	**0.704**	0.292	**0.709**	0.246
X+NCC	0.607	0.647	**0.423**	0.656	0.342	0.648	**0.343**

a clearly distinguishable object is equivalent to the addition of label noise which is known to hurt robustness [27]. In this section we show that the addition of samples with perturbations $\delta \geq r'_j$ for $r'_j < 0.5r_j$ indeed hurts the performance of classifiers according to several metrics on CIFAR-10. Due to space constraints we present further affirmative results for SVHN in Appendix F.

Extension by the globally critical points As $|\{\bar{x}\}^{\text{crit}}_{\text{global}}| = 132$ for CIFAR-10 (see Table 1), their addition is not measurably impacting generalization performance (see Appendix H). Thus, to simulate different levels of label noise we add random samples from the original train set $X^{\text{org}}_{\text{train}}$ to $\{\bar{x}\}^{\text{crit}}_{\text{global}}$ to obtain train sets with different relative amounts of original and critical samples and therefore different amounts of label noise. This experimental setup roughly follows Sanyal et al. [27]. We train a neural network on these datasets and measure its accuracy on the original train and test set and against FGSM attacks [13] of different strengths.

In Figure 5a we observe that with increasing label noise test accuracy deteriorates while adversarial accuracy against FGSM attacks increases. Although, due to the small train set, test accuracy is already low, adding samples with no visible class object further deteriorates test accuracy as the model is likely biased towards learning superficial textural clues. Train accuracy on the other hand is not hurt, as those samples can simply be memorized. As $\{\bar{x}\}^{\text{crit}}_{\text{global}}$ are defined by having $r'_j \leq 0.5R$, the distance between $\bar{x}'_j$ and x'_j is small and thus small-norm perturbations as those introduced by FGSM do not result in wrong predictions as the network interpolates between $\bar{x}'_j$ and x'_j.

Extension by the locally critical points It is common practice in adversarial training to pick a single perturbation magnitude δ for all samples under the assumption that no class change is induced by its application. However, this procedure is suboptimal and error-prone as upper-bounds on δ can be influenced by labelling errors in the original train set. Thus, more recent robust training methods work with instance-specific perturbation magnitudes [7].

To further show that the addition of samples for which $\delta \geq R^*$ leads to reduced generalization performance, we create two new train sets by appending either $\{\bar{x}\}^{\text{crit}}_{\text{local}}$ with the label of its corresponding x'_j (no class change, NCC), denoted X+NCC, or the label of its corresponding nearest neighbours $\{x_i\}^d_{i=1}$ (class change, CC), denoted X+CC, to the original train set $X^{\text{org}}_{\text{train}}$. We train a network on each of these three datasets and report accuracies on the original train and original test set in Figures 5b and 5c. We observe that while train accuracy is not hurt, test accuracy deteriorates when trained on X+NCC. This is likely due to $\{\bar{x}\}^{\text{crit}}_{\text{local}}$ biasing the model towards learning superficial textural clues which does not deteriorate train but does reduce generalization performance. The addition of $\{\bar{x}\}^{\text{crit}}_{\text{local}}$ when assigned a different class label than x'_j improves generalization performance. This is likely due to the network interpolating between $\{x_i\}^d_{i=1}$ which appears to help for CIFAR-10. Contrary, in Appendix F we show for SVHN that both train sets X+NCC and X+CC reduce generalization performance.

Finally, we also test these models against a benchmark of common perturbations [14]. We obtain similar results to the label noise experiments above. In Table 5 we observe that accuracy against small-norm noise perturbations is increased whereas accuracy against large-norm blur perturbations is mostly decreased. Intuitively, a flat loss surface around training points or obfuscated gradients [74] help to protect against small-norm changes, whereas large-norm changes need to be defend against by learning semantic concepts.

These results show that the maximum perturbation magnitudes for robust training need to be chosen carefully as they can deteriorate the generalisation to accuracy and robustness benchmarks while train accuracy is unharmed.

6 Conclusions

Robustness and generalization behaviour of neural networks have traditionally been studied by investigating properties of their learned representation or their training methods. More recently, properties of datasets came into focus as potential causes for their generalization and robustness deficits (see Section 2). This work contributes to this line of work. We study the complexity of decision boundaries for robust training in a model-agnostic way and derive a lower bound on the perturbation magnitude that increase their complexity. For common image benchmarks it also bounds the introduction of label noise which we show to hurt generalization and robustness. Thus, our work shows that studying geometric properties of data distributions can yield practical insights into modern deep classifiers and can provide guidelines for the choice of architectures and training parameters.

Acknowledgements D. Kienitz and E. Komendantskaya acknowledge support of EPSRC grant EP/T026952/1: *AI Secure and Explainable by Construction (AISEC).*

References

1. He, W., Li, B., Song, D.: Decision boundary analysis of adversarial examples. In: International Conference on Learning Representations. (2018)
2. Fawzi, A., Moosavi-Dezfooli, S.M., Frossard, P., Soatto, S.: Classification regions of deep neural networks. arXiv preprint arXiv:1705.09552 (2017)
3. Ortiz-Jimenez, G., Modas, A., Moosavi, S.M., Frossard, P.: Hold me tight! influence of discriminative features on deep network boundaries. Advances in Neural Information Processing Systems **33** (2020) 2935–2946
4. Narayanan, H., Mitter, S.: Sample complexity of testing the manifold hypothesis. In: Advances in neural information processing systems. (2010) 1786–1794
5. Narayanan, H., Niyogi, P.: On the sample complexity of learning smooth cuts on a manifold. In: COLT. (2009)
6. Kienitz, D., Komendantskaya, E., Lones, M.: The effect of manifold entanglement and intrinsic dimensionality on learning. In: 36th AAAI Conference on Artificial Intelligence 2022, AAAI Press (2021)
7. Ding, G.W., Sharma, Y., Lui, K.Y.C., Huang, R.: Mma training: Direct input space margin maximization through adversarial training. In: International Conference on Learning Representations. (2019)
8. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. (2016) 1135–1144
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
10. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
11. Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F.A., Brendel, W.: Partial success in closing the gap between human and machine vision. Advances in Neural Information Processing Systems **34** (2021) 23885–23899
12. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations, ICLR 2014. (2014)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
14. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations. (2018)
15. Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L.: Measuring robustness to natural distribution shifts in image classification. Advances in Neural Information Processing Systems **33** (2020) 18583–18599
16. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International Conference on Machine Learning, PMLR (2019) 5389–5400
17. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations. (2018)
18. Raghunathan, A., Xie, S.M., Yang, F., Duchi, J.C., Liang, P.: Adversarial training can hurt generalization. arXiv preprint arXiv:1906.06032 (2019)

19. Zhang, X., Chen, J., Gu, Q., Evans, D.: Understanding the intrinsic robustness of image distributions using conditional generative models. In: International Conference on Artificial Intelligence and Statistics, PMLR (2020) 3883–3893
20. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152 (2018)
21. Stutz, D., Hein, M., Schiele, B.: Disentangling adversarial robustness and generalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 6976–6987
22. Yang, Y.Y., Rashtchian, C., Wang, Y., Chaudhuri, K.: Robustness for non-parametric classification: A generic attack and defense. In: International Conference on Artificial Intelligence and Statistics, PMLR (2020) 941–951
23. Shah, H., Tamuly, K., Raghunathan, A., Jain, P., Netrapalli, P.: The pitfalls of simplicity bias in neural networks. Advances in Neural Information Processing Systems **33** (2020) 9573–9585
24. Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., Madry, A.: Adversarially robust generalization requires more data. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. (2018) 5019–5031
25. Yin, D., Kannan, R., Bartlett, P.: Rademacher complexity for adversarially robust generalization. In: International conference on machine learning, PMLR (2019) 7085–7094
26. Nakkiran, P.: Adversarial robustness may be at odds with simplicity. arXiv preprint arXiv:1901.00532 (2019)
27. Sanyal, A., Dokania, P.K., Kanade, V., Torr, P.: How benign is benign overfitting? In: International Conference on Learning Representations. (2020)
28. Nguyen, Q., Mukkamala, M.C., Hein, M.: Neural networks should be wide enough to learn disconnected decision regions. In: International Conference on Machine Learning, PMLR (2018) 3740–3749
29. Yang, Y.Y., Rashtchian, C., Zhang, H., Salakhutdinov, R.R., Chaudhuri, K.: A closer look at accuracy vs. robustness. Advances in neural information processing systems **33** (2020) 8588–8601
30. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. Advances in neural information processing systems **32** (2019)
31. Joe, B., Hwang, S.J., Shin, I.: Learning to disentangle robust and vulnerable features for adversarial detection. arXiv preprint arXiv:1909.04311 (2019)
32. Singla, S., Feizi, S.: Salient imagenet: How to discover spurious features in deep learning? In: International Conference on Learning Representations. (2021)
33. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 427–436
34. Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org (2017) 233–242
35. Hermann, K., Lampinen, A.: What shapes feature representations? exploring datasets, architectures, and training. Advances in Neural Information Processing Systems **33** (2020) 9995–10006
36. Ahmed, F., Bengio, Y., van Seijen, H., Courville, A.: Systematic generalisation with group invariant predictions. In: International Conference on Learning Representations. (2020)

37. Valle-Perez, G., Camargo, C.Q., Louis, A.A.: Deep learning generalizes because the parameter-function map is biased towards simple functions. In: International Conference on Learning Representations. (2018)
38. Jo, J., Bengio, Y.: Measuring the tendency of cnns to learn surface statistical regularities. arXiv preprint arXiv:1711.11561 (2017)
39. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European conference on computer vision (ECCV). (2018) 456–473
40. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations. (2018)
41. Geirhos, R., Medina Temme, C., Rauber, J., Schütt, H., Bethge, M., Wichmann, F.: Generalisation in humans and deep neural networks. In: Thirty-second Annual Conference on Neural Information Processing Systems 2018 (NeurIPS 2018), Curran (2019) 7549–7561
42. Hermann, K., Chen, T., Kornblith, S.: The origins and prevalence of texture bias in convolutional neural networks. Advances in Neural Information Processing Systems **33** (2020) 19000–19015
43. Carter, B., Jain, S., Mueller, J.W., Gifford, D.: Overinterpretation reveals image classification model pathologies. Advances in Neural Information Processing Systems **34** (2021)
44. Singla, S., Nushi, B., Shah, S., Kamar, E., Horvitz, E.: Understanding failures of deep networks via robust feature extraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 12853–12862
45. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35** (2013) 1798–1828
46. Bhagoji, A.N., Cullina, D., Mittal, P.: Lower bounds on adversarial robustness from optimal transport. Advances in Neural Information Processing Systems **32** (2019) 7498–7510
47. Dobriban, E., Hassani, H., Hong, D., Robey, A.: Provable tradeoffs in adversarially robust classification. arXiv preprint arXiv:2006.05161 (2020)
48. Dan, C., Wei, Y., Ravikumar, P.: Sharp statistical guaratees for adversarially robust gaussian classification. In: International Conference on Machine Learning, PMLR (2020) 2345–2355
49. Bhattacharjee, R., Jha, S., Chaudhuri, K.: Sample complexity of robust linear classification on separated data. In: International Conference on Machine Learning, PMLR (2021) 884–893
50. Khim, J., Loh, P.L.: Adversarial risk bounds via function transformation. arXiv preprint arXiv:1810.09519 (2018)
51. Attias, I., Kontorovich, A., Mansour, Y.: Improved generalization bounds for robust learning. In: Algorithmic Learning Theory, PMLR (2019) 162–183
52. Montasser, O., Hanneke, S., Srebro, N.: Vc classes are adversarially robustly learnable, but only improperly. In: Conference on Learning Theory, PMLR (2019) 2512–2530
53. Ashtiani, H., Pathak, V., Urner, R.: Black-box certification and learning under adversarial perturbations. In: International Conference on Machine Learning, PMLR (2020) 388–398
54. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical

analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 8340–8349
55. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. In: International Conference on Learning Representations. (2019)
56. Rebuffi, S.A., Gowal, S., Calian, D.A., Stimberg, F., Wiles, O., Mann, T.A.: Data augmentation can improve robustness. Advances in Neural Information Processing Systems **34** (2021)
57. Hendrycks, D., Lee, K., Mazeika, M.: Using pre-training can improve model robustness and uncertainty. arXiv preprint arXiv:1901.09960 (2019)
58. Carmon, Y., Raghunathan, A., Schmidt, L., Liang, P., Duchi, J.C.: Unlabeled data improves adversarial robustness. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. (2019) 11192–11203
59. Alayrac, J.B., Uesato, J., Huang, P.S., Fawzi, A., Stanforth, R., Kohli, P.: Are labels required for improving adversarial robustness? Advances in Neural Information Processing Systems **32** (2019) 12214–12223
60. Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., Kohli, P.: Adversarial robustness through local linearization. Advances in Neural Information Processing Systems **32** (2019)
61. Ross, A., Doshi-Velez, F.: Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 32. (2018)
62. Chan, A., Tay, Y., Ong, Y.S., Fu, J.: Jacobian adversarially regularized networks for robustness. In: International Conference on Learning Representations. (2020)
63. Etmann, C., Lunz, S., Maass, P., Schönlieb, C.: On the connection between adversarial robustness and saliency map interpretability. In: ICML. (2019)
64. Simpson, B., Dutil, F., Bengio, Y., Cohen, J.P.: Gradmask: Reduce overfitting by regularizing saliency. In: International Conference on Medical Imaging with Deep Learning–Extended Abstract Track. (2019)
65. Fawzi, A., Moosavi-Dezfooli, S.M., Frossard, P., Soatto, S.: Empirical study of the topology and geometry of deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 3762–3770
66. Andriushchenko, M., Flammarion, N.: Understanding and improving fast adversarial training. Advances in Neural Information Processing Systems **33** (2020) 16048–16059
67. Rice, L., Wong, E., Kolter, Z.: Overfitting in adversarially robust deep learning. In: International Conference on Machine Learning, PMLR (2020) 8093–8104
68. LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: Advances in neural information processing systems. (1990) 396–404
69. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
70. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011. (2011)
71. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)
72. Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., Hein, M.: Robustbench: a standardized adversarial robustness benchmark. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track. (2021)

73. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (Poster). (2015)
74. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. arXiv preprint arXiv:1802.00420 (2018)
75. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. (1992) 144–152
76. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2574–2582
77. Addepalli, S., Jain, S., Sriramanan, G., Khare, S., Radhakrishnan, V.B.: Towards achieving adversarial robustness beyond perceptual limits. In: ICML 2021 Workshop on Adversarial Machine Learning. (2021)
78. Augustin, M., Meinke, A., Hein, M.: Adversarial robustness on in-and out-distribution improves explainability. In: European Conference on Computer Vision, Springer (2020) 228–245
79. Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., Tsipras, D.: Robustness (python library) (2019)
80. Kireev, K., Andriushchenko, M., Flammarion, N.: On the effectiveness of adversarial training against common corruptions. In: Uncertainty in Artificial Intelligence, PMLR (2022) 1012–1021
81. Modas, A., Rade, R., Ortiz-Jiménez, G., Moosavi-Dezfooli, S.M., Frossard, P.: Prime: A few primitives can boost robustness to common corruptions. arXiv preprint arXiv:2112.13547 (2021)
82. Rade: Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In: ICML 2021 Workshop on Adversarial Machine Learning. (2021)
83. Wong, E., Rice, L., Kolter, J.Z.: Fast is better than free: Revisiting adversarial training. In: International Conference on Learning Representations. (2019)
84. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, PMLR (2015) 448–456
85. Fukushima, K.: Visual feature extraction by a multilayered network of analog threshold elements. IEEE Transactions on Systems Science and Cybernetics **5** (1969) 322–333
86. Fukushima, K., Miyake, S.: Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: Competition and cooperation in neural nets. Springer (1982) 267–285
87. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings (2011) 315–323
88. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32. Curran Associates, Inc. (2019) 8024–8035