ADVERSARIAL EXAMPLES ARE A NATURAL CONSEQUENCE OF TEST ERROR IN NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Maliciously constructed inputs, or *adversarial examples*, can fool trained machine learning models. Over the last few years, adversarial examples have captured the attention of the research community, especially in the case where the adversary is restricted to making only small modifications of a correctly handled input. When it was first discovered that neural networks are sensitive to small perturbations, many researchers found this surprising and proposed several hypotheses to explain it. In this work, we show that this sensitivity and the poor performance of classification models (relative to humans) on noisy images are two manifestations of the same underlying phenomenon. Nearby errors simply lie on the boundary of a large set of errors whose volume can be measured using test error in additive noise. We present compelling new evidence in favor of this interpretation before discussing some preexisting results which also support our perspective. The relationship between nearby errors and failure to generalize in noise has implications for the adversarial defense literature, as it suggests that defenses which fail to reduce test error in noise will also fail to defend against small adversarial perturbations. This yields a computationally tractable evaluation metric for defenses to consider: test error in noisy image distributions.

1 INTRODUCTION

Since the publication of Szegedy et al. (2014), a large body of literature has arisen in an attempt to address the fact that statistical classifiers are sensitive to small perturbations of the input, so-called "adversarial examples." Even when, for example, an image is classified correctly with high confidence, it is often the case that a very nearby image is classified incorrectly. An assumption underlying most of this work is that solving this problem requires a different set of methods than the ones being developed to improve model generalization. The adversarial defense literature focuses primarily on improving robustness to small perturbations of the input and rarely reports improved generalization in *any* distribution.

This perspective is encapsulated by the following question from Szegedy et al. (2014): "The existence of the adversarial negatives appears to be in contradiction with the network's ability to achieve high generalization performance. Indeed, if the network can generalize well, how can it be confused by these adversarial negatives, which are indistinguishable from the regular examples?"

To understand this apparent contradiction, though, it is important to recognize that the inputs to these image classifications models live in a very high-dimensional space whose geometry can be counterintuitive. This paper presents evidence that the existence of these nearby errors can be easily explained by basic facts about high-dimensional geometry and is not indicative of any kind of "flaw" in model architecture that is distinct from imperfect generalization performance. In particular, given the error rates we observe in *noisy* image distributions, the existence of adversarial examples should be expected.

Understanding the relationship between nearby errors and model generalization requires understanding the geometry of the error set of a statistical classifier. In particular, the assertion that these adversarial examples are a distinct phenomenon from test error is equivalent to stating that the error set is in some sense poorly behaved. In order to formalize this notion, we study two functions of a model's error set E:

- The first quantity, *test error* or *generalization error* under a given distribution of inputs $q(x)$, is the probability that a random sample from the distribution q is in E. We will denote this $\mathbb{P}_{x\sim q}[x \in E]$; reducing this quantity is the goal of supervised learning. Note that q need not be restricted to the distribution from which the training set was sampled.
- The second quantity is called *adversarial robustness*. For an input x and a metric on the input space d, let $d(x, E)$ denote the distance from x to the nearest point of E. For any ϵ, let E_ϵ denote the set $\{x : d(x, E) < \epsilon\}$, the set of points within ϵ of an error. The adversarial robustness of the model is then $\mathbb{P}_{x\sim q}[x \in E_\epsilon]$, the probability that a random sample from q is within distance ϵ of some point in the error set. Reducing this quantity is the goal of much of the adversarial defense literature. When we refer to "adversarial examples" in this paper, we will always mean these nearby errors.

In geometric terms we can think of $\mathbb{P}_{x\sim q}[x \in E]$ as a sort of volume of the error set while $\mathbb{P}_{x\sim q}[x \in E_\epsilon]$ is related to its surface area. More directly, $\mathbb{P}_{x\sim q}[x \in E_\epsilon]$ is what we will call the *ϵ-boundary measure*, the volume under q of the region within ϵ of the surface or the interior. For example, in each image in Figure 1, if E is the pink region, E_ϵ is the union of the pink and blue regions.

The phenomenon that surprised Szegedy et al. (2014) is then simply that, for small ϵ, $\mathbb{P}_{x\sim q}[x \in E_\epsilon]$ can be large even when $\mathbb{P}_{x\sim q}[x \in E]$ is small. In other words, most correctly classified inputs are very close to a misclassified point, even though the model is very accurate. This phenomenon is ubiquitous across machine learning model classes and datasets — it is not specific to neural networks (Papernot et al., 2017). Despite a large amount of effort, though, there is currently no known general solution. There have been some specific metrics and ϵ's for which adversarial robustness has been successfully increased (Madry et al., 2017), but these models remain sensitive to small perturbations in other metrics (Sharma & Chen, 2017), or slightly larger ϵ.

In high-dimensional spaces this phenomenon is not isolated to the error sets of statistical classifiers, but in fact *almost every* nonempty set of small volume has large ϵ-boundary measure, even sets that seem very well-behaved. As a simple example, consider the measure of the set $E = \{x \in \mathbb{R}^n : ||x||_2 < 1\}$ under the Gaussian distribution $q = \mathcal{N}(0, \sigma^2 I)$. For $n = 1000$, $\sigma = 1.05/\sqrt{n}$, and $\epsilon = 0.1$, we have $\mathbb{P}_{x\sim q}[x \in E] \approx 0.02$ and $\mathbb{P}_{x\sim q}[x \in E_\epsilon] \approx 0.98$, so most samples from q will be close to E despite the fact that E has relatively little measure under the Gaussian distribution. A model with this error set would be very accurate but still sensitive to small adversarial perturbations of the input. If we relied only on our low-dimensional spatial intuition, we might be surprised to find how consistently these small adversarial perturbations could be found — 98% of our test points would have an error at distance 0.1 or less even though only 2% are misclassified. However, taking this very local view of our model's decision boundary would prevent us from seeing the larger picture: that these seemingly surprising errors were in fact just the nearest points in the same large region of input space whose volume we estimated when computing the test error. Additional examples of the relationship between volume and surface area in high dimensions, but where q is a uniform measure rather than a Gaussian, appear in Figure 1.

In the example in the previous paragraph, the fact that $\mathbb{P}_{x\sim q}[x \in E_\epsilon]$ is so large even though $\mathbb{P}_{x\sim q}[x \in E]$ is small is solely a consequence of the large dimension of the space rather than any pathological properties of E. (E is in fact just a ball!) It is difficult to use two- or three-dimensional visual intuition to reason about the relationship between these two quantities; constructing an example in two dimensions with the same $\mathbb{P}_{x\sim q}[x \in E]$ and $\mathbb{P}_{x\sim q}[x \in E_\epsilon]$ requires a much more complicated set.

The discussion of high-dimensional geometry suggests that adversarial examples may actually not be in contradiction to high generalization performance. Indeed, high generalization performance does not mean *perfect* generalization, and in high-dimensional spaces even a small amount of test error can lie close to most correctly classified inputs. In this paper, we explore the connection between small perturbation adversarial examples and generalization in noise in several different ways.

First, in Section 4, we establish a relationship between the error rate of an image classification model in the presence of Gaussian noise and the existence of adversarial examples for *noisy* versions of test set images. In this setting we can actually prove a rigorous, model-independent bound relating these two quantities that is achieved when the error set is a half space, and we see that the models we tested are already quite close to this optimum. Therefore, for these noisy image distributions the only way to defend against adversarial examples is to improve generalization.

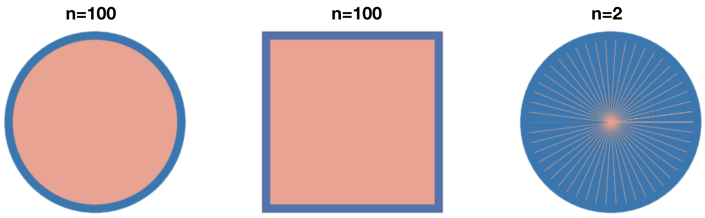

Figure 1: The relationship between volume and surface area behaves very counterintuitively in high dimensions. In the first two images the distribution q is uniform over a 100-dimensional ball (blue region, left) or a cube (blue region, center) and the error set E, in pink, is a smaller ball or cube. We see that allowing the error set to be even slightly smaller makes its volume a tiny fraction of the total even though every point is still very close to an error.
In the left image, the blue ball has radius 1.1, and the pink ball has radius 1.0, so $\mathbb{P}_{x\sim q}[x \in E] = (1.0/1.01)^{100} = 7.3 \times 10^{-5}$ and $\mathbb{P}_{x\sim q}[x \in E_\epsilon] = 1.0$ when $\epsilon = .1$.
In the center image, the blue cube has side length 2.2, and the pink cube has side length 2.0, so $\mathbb{P}_{x\sim q}[x \in E] = (2.0/2.2)^{100} = 7.3 \times 10^{-5}$ and $\mathbb{P}_{x\sim q}[x \in E_\epsilon] = 1.0$ when $\epsilon = .1$.
Although the volume of the pink region appears large in these diagrams, this is misleading: we are only looking at two-dimensional slices of 100-dimensional objects. A similar relationship between volume and surface area in two dimensions requires a much stranger-looking error set, as in the rightmost image where $\mathbb{P}_{x\sim q}[x \in E] = .15$ and $\mathbb{P}_{x\sim q}[x \in E_\epsilon] = 1.0$.

Next, in Section 5, we relate generalization in Gaussian noise to adversarial perturbations of the *clean* image. Here it is not possible to prove a theorem as strong as the one in Section 4. Still, we provide evidence that the errors we find close to the clean image and the errors we sample under Gaussian noise are part of the same large set. Furthermore, the existence of the nearby error is due mostly to the large dimension of the input space rather than any strangeness in the shape of the error set. This analysis builds upon prior work (Fawzi et al., 2018; 2016) which makes smoothness assumptions on the decision boundary to relate these two quantities.

Finally, this relationship suggests that training procedures designed to improve adversarial robustness might improve generalization in the presence of noise and vice versa. In Section 6, we present evidence that this is indeed the case.

2 Related Work

The broader field of *adversarial machine learning* researches general ways in which an adversary may interact with an ML system, and dates back to 2004 (Dalvi et al., 2004; Biggio & Roli, 2018). Since the work of Szegedy et al. (2014), a subfield has focused specifically on the phenomenon of small adversarial perturbations of the input, or "adversarial examples." In Szegedy et al. (2014) it was proposed these adversarial examples occupy a dense, measure-zero subset of image space. However, more recent work has provided evidence that this is not true. For example, Fawzi et al. (2016); Franceschi et al. (2018) shows that under linearity assumptions of the decision boundary small adversarial perturbations exist when test error in noise is non-zero. Gilmer et al. (2018b) showed for a specific data distribution that there is a fundamental upper bound on adversarial robustness in terms of test error. Mahloujifar et al. (2018) has generalized these results to a much broader class of distributions.

Recent work has proven for a synthetic data distribution that adversarially robust generalization requires more data (Schmidt et al., 2018). The distribution they consider when proving this result is a mixture of high dimensional Gaussians. As we will soon discuss, every set E of small measure in the high dimensional Gaussian distribution has large boundary measure. Therefore, at least for the data distribution considered, the main conclusion of this work is that "adversarially robust generalization requires more data" is a direct corollary of the statement "generalization requires more data."

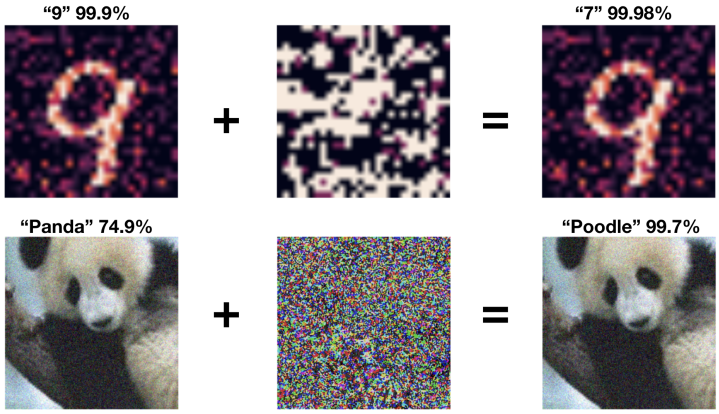

Figure 2: The adversarial example phenomenon occurs for noisy images as well as clean ones. Starting with a noisy image that that is correctly classified, one can apply carefully crafted imperceptible noise to it which causes the model to output an incorrect answer. This occurs even though the error rate among random Gaussian perturbations of this image is small (less than $.1\%$ for both the MNIST 9 and the ImageNet panda shown above). In fact, we prove that the presence of errors in Gaussian noise *logically implies* that small adversarial perturbations exists around noisy images. The only way to "defend" against such adversarial perturbations is to reduce the error rate in Gaussian noise.

3 Models Considered

In this work we will investigate several different models trained on the MNIST, CIFAR-10 and ImageNet datasets. For MNIST and CIFAR-10 we look at the naturally trained and adversarially trained models which have been open-sourced by Madry et al. (2017). We also trained the same model on CIFAR-10 with Gaussian data augmentation. For ImageNet, we investigate Wide ResNet-50 trained with Gaussian data augmentation. Unfortunately, we were unable to study the effects of adversarial training on ImageNet because no robust open sourced model exists (we considered the models released in Tramèr et al. (2017) but found that they only minimally improve robustness to the white box PGD adversaries we consider here). Additional training details can be found in Appendix A.

4 Errors in Noise Imply Adversarial Examples for Noisy Images

The Gaussian Isoperimetric Inequality. Let x be a correctly classified image and consider the distribution q of Gaussian perturbations of x with some fixed variance $\sigma^2 I$. For this distribution, there is a precise sense in which small adversarial perturbations exist only because test error is nonzero. That is, given the error rates we actually observe on noisy images, most noisy images *must* be close to the error set. This result holds completely independently of any assumptions about the model and follows from a fundamental geometric property of the high-dimensional Gaussian distribution, which we will now make precise.

For an image x and the corresponding noisy image distribution q, let $\epsilon_q^*(E)$ be the median distance from one of these noisy images to the nearest error. (In other words, it is the ϵ for which $\mathbb{P}_{x\sim q}[x \in E_\epsilon] = \frac{1}{2}$.) As before, let $\mathbb{P}_{x\sim q}[x \in E]$ be the probability that a random Gaussian perturbation of x lies in E. It is possible to deduce a bound relating these two quantities from the *Gaussian isoperimetric inequality* (Borell, 1975). The form we will use is:

Theorem (Gaussian Isoperimetric Inequality)**.** *Let $q = \mathcal{N}(0, \sigma^2 I)$ be the Gaussian distribution on $\mathbb{R}^n$ with variance $\sigma^2 I$, and let $\mu = \mathbb{P}_{x\sim q}[x \in E]$.*

Write $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp(-x^2/2)dx$, the cdf of the univariate standard normal distribution. If $\mu \geq \frac{1}{2}$, then $\epsilon_q^*(E) = 0$. Otherwise, $\epsilon_q^*(E) \leq -\sigma\Phi^{-1}(\mu)$, with equality when E is a half space.

In particular, for any machine learning model for which the error rate in the distribution q is at least μ, the median distance to the nearest error is at most $-\sigma\Phi^{-1}(\mu)$. (Note that $\Phi^{-1}(\mu)$ is negative when $\mu < \frac{1}{2}$.) Because each coordinate of a multivariate normal is a univariate normal, $-\Phi^{-1}(\mu)$ is the distance to a half space for which the error rate is μ when $\sigma = 1$.

Note that the isoperimetric bound depends only on the error rate μ and the standard deviation σ of a single component, and not directly on the dimension. This might seem at odds with the emphasis on high-dimensional geometry in Section 1. The dimension does appear if we consider the norm of a typical sample from $\mathcal{N}(0, \sigma^2 I)$, which is $\sigma\sqrt{n}$. As the dimension increases, so does the ratio between the distance to a noisy image and the distance at which we expect to find the decision boundary.

In Appendix E we will give the more common statement of the Gaussian isoperimetric inequality along with a proof of the version presented here. In geometric terms, we can say that a half space is the set E of a fixed volume that minimizes the surface area under the Gaussian measure, similar to how a circle is the set of fixed area that minimizes the perimeter. So among models with some fixed test error $\mathbb{P}_{x\sim q}[x \in E]$, the most robust on this distribution are the ones whose error set is a half space.

Comparing Neural Networks to the Isoperimetric Bound. We evaluated these quantities for several models and many images from the CIFAR-10 and ImageNet test sets. Just like for clean images, we found that most noisy images are both correctly classified and very close to a visually similar image which is not. (See Figure 2.)

It is not actually possible to compute ϵ_q^* precisely for the error set of a neural network. In fact, finding the distance to the nearest error is NP-hard (Katz et al., 2017). Instead, the best we can do is to search for an error using a method like PGD (Madry et al., 2017) and report the nearest error we can find. For each test image, we took 1,000 samples from the corresponding Gaussian and estimated ϵ_q^* using PGD with 200 steps on each sample and reported the median.

We find that for the five models we considered on CIFAR-10 and ImageNet, the relationship between our estimate of $\epsilon_q^*(E)$ and $\mathbb{P}_{x\sim q}[x \in E]$ is already close to optimal. This is visualized in Figure 3. Note that in both cases, adversarial training does improve robustness to small perturbations, but the gains are primarily because error rates in Gaussian noise were dramatically improved, and less because the surface area of the error set was decreased. In particular, many test points do not appear on these graphs because error rates in noise were so low that we did not find any errors among the 100,000 samples we used. For example, for the naturally trained CIFAR model, about 1% of the points lie off the left edge of the plot, compared to about 59% for the adversarially trained model and 70% for the model trained on noise. This shows that adversarial training on small perturbations improved generalization to large random perturbations, as the isoperimetric inequality says it must.

Not all models or functions will be this close to optimal. As a simple example, if we took one of the CIFAR models shown in Figure 3 and modified it so that the model outputs an error whenever each coordinate of the input is an integer multiple of 10^{-6}, the resulting model would have an error within $\sqrt{\frac{1}{2} \cdot 10^{-6} \cdot \dim(\text{CIFAR})} \approx 0.039$ of every point. In this case, adversarial examples *would* be a distinct phenomenon from generalization, since $\epsilon_q^*(E)$ would be far from optimal.

The contrast between these two settings is important for adversarial defense design. If adversarial examples arose from a badly behaved decision boundary (as in the latter case), then it would make sense to design defenses which attempt to smooth out the decision boundary in some way. However, because we observe that image models are already close to the optimal bound on robustness for a fixed error rate in noise, future defense design should attempt to improve generalization in noise. Currently there is a considerable subset of the adversarial defense literature which develops methods that would remove any small "pockets" of errors but which don't improve model generalization. (This is discussed further in Section 6.) One example is Xie et al. (2017) which proposes randomly resizing the input to the network as a defense strategy. Unfortunately, this defense, like many others, has been shown to be ineffective against stronger adversaries (Carlini & Wagner, 2017a;b; Athalye et al., 2018).

Note that moving a point directly upward in Figure 3 amounts to rearranging the error set so that it has the same volume but a smaller surface area. This does not improve generalization, nor does it make it harder for an adversary to find an error. Furthermore, even if there is a way to rearrange

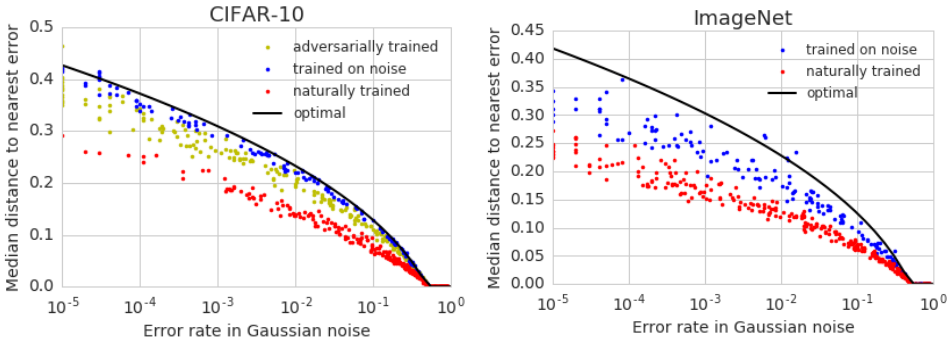

Figure 3: Plots of the error rate in Gaussian noise vs. median distance from a noisy point to the nearest error (found using PGD) for the CIFAR-10 and ImageNet models we examined. Both plots used noise with $\sigma = 0.1$. Each dot corresponds to one test image, and the black line shows the bound from the isoperimetric inequality, which is achieved when the error set is a half space. Both plots include data from a model trained on clean images, an adversarially trained model, and a model trained on Gaussian noise ($\sigma = 0.4$.) As mentioned in Section 3, we were unable to run this experiment on an adversarially robust ImageNet model.

errors to reduce the surface area, it is unclear what purpose this would serve unless the goal is *only* to eliminate errors in some small l_p ball.

5 Errors in Noise Suggest Adversarial Examples for Clean Images

The Linear Case. The isoperimetric inequality does not say anything about errors close to the clean image x; it only says that most samples from $\mathcal{N}(x, \sigma^2 I)$ must be close to an error. In fact, it is possible to construct error sets for which the decision boundary is far away but there are many errors in noise.

However, for linear models, the relationship between errors in Gaussian noise and small perturbations of the *clean* image is exact. For a clean image x, let $d(x)$ be the distance from x to decision boundary, and let $\sigma(x, \mu)$ be the σ for which $\mathbb{P}_{x \sim q}[x \in E]$ is some fixed error rate μ. The relationship between these two quantities is especially easy to write down: we just have $d = -\sigma(x, \mu)\Phi^{-1}(\mu)$. (We have the same indirect dependence on dimension here as we saw in Section 4: the distance to a typical sample from the Gaussian is $\sigma\sqrt{n}$.)

Experiments suggest that the small-scale behavior of neural networks is close enough to linear to explain the adversarial perturbations of the clean image we can find. This relationship was explored in detail in Fawzi et al. (2016; 2018). Here we present some additional empirical evidence.

We examined the relationship between d and σ for neural networks when $\mu = 0.01$. For each test point, we compared $\sigma(x, \mu)$ to an estimate of $d(x)$ computed using PGD. The results for both the ordinarily and adversarially trained CIFAR-10 models are shown in Figure 4, together with a line representing how these quantities would be related for a linear model. As we saw in Section 4, adversarial training succeeds in increasing $d(x)$, but it does so while also increasing $\sigma(x, \mu)$.

While the adversarially trained model deviates more from the linear case than the naturally trained model, it does so in the direction of *greater* distances to the decision boundary. So, to explain the distances to the errors we can find using PGD, it is not necessary to rely on any great complexity in the shape of the error set; a linear model with the same error rates in noise would have errors just as close.

Visualizing the Decision Boundary. In Figure 5 we drew some pictures of two-dimensional slices of image space through several different triples of points. (Similar visualizations have previously appeared in Fawzi et al. (2018), and are called "church window plots.")

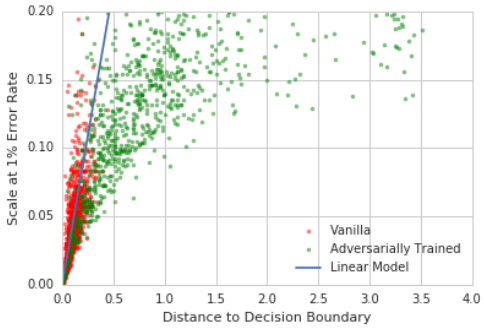

Figure 4: Comparing the distance to decision boundary with the σ for which the error rate in Gaussian noise is 1%. We see both the vanilla and adversarially trained models behave like a linear model at smaller scales, where the connection between test error and adversarial examples is most direct.

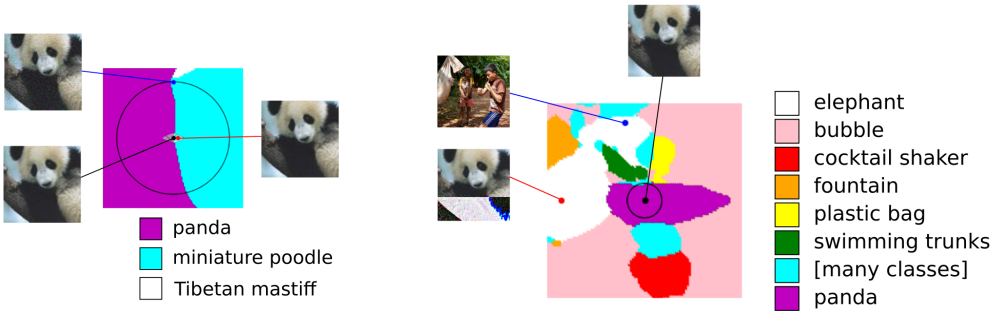

Figure 5: Two-dimensional slices of image space through different triples of points together with the classes assigned by a trained model. The black circle in both images has radius 31.4, corresponding to noise with $\sigma = 31.4/\sqrt{n} = 0.08$.
Left: An image from the test set (black), a random misclassified Gaussian perturbation at standard deviation 0.08 (blue), and an error found using PGD (red). The estimated measure of the cyan region ("miniature poodle") in the Gaussian distribution is about 0.1%. The small diamond-shaped region in the center of the image is the l_∞ ball of radius 8/255.
Right: A slice at a larger scale with the same black point, together with an error from the clean set (blue) and an adversarially constructed error (red) which are both assigned to the same class ("elephant").

We see some common themes. In the figure on the left, we see that an error found in Gaussian noise lies in the same connected component of the error set as an error found using PGD, and that at this scale that component visually resembles a half space. This figure also illustrates the relationship between test error and adversarial robustness. To measure adversarial robustness is to ask whether or not there are any errors in the l_∞ ball — the small diamond-shaped region in the center of the image — and to measure test error in noise is to measure the volume of the error set in the defined noise distribution. At least in this slice, nothing distinguishes the PGD error from any other point in the error set apart from its proximity to the center point.

The figure on the right shows a different slice through the same test point but at a larger scale. This slice includes an ordinary test error along with an adversarial perturbation of the center image constructed with the goal of maintaining visual similarity while having a large l_2 distance. The two errors are both classified (incorrectly) by the model as "elephant." This adversarial error is actually *farther* from the center than the test error, but they still clearly belong to the same connected component. This suggests that defending against worst-case content-preserving perturbations (Gilmer et al., 2018a) requires removing all errors at a scale comparable to the distance between unrelated pairs of images. Many more church window plots can be found in Appendix F.

Dataset	CIFAR-10				ImageNet		
Training	Vanilla	Noise $\sigma = 0.1$	Noise $\sigma = 0.4$	Adv	Vanilla	Noise $\sigma = 0.4$	Noise $\sigma = 0.8$
Noise Type							
Clean	95.0%	93.5%	84.0%	87.3%	76.0%	74.4%	72.6%
PCA100, $\sigma = 0.2$	93.2%	92.3%	83.6%	86.5%	45.5%	56.5%	59.7%
PCA100, $\sigma = 0.4$	82.6%	83.1%	81.0%	80.6%	13.5%	17.7%	19.7%
Pepper, $p = 0.1$	20.2%	53.3%	81.2%	38.4%	31.3%	70.0%	69.1%
Pepper, $p = 0.3$	12.3%	18.9%	58.0%	21.1%	5.4%	56.0%	61.5%
Gaussian, $\sigma = 0.1$	29.1%	89.0%	85.1%	77.8%	60.7%	73.3%	71.7%
Gaussian, $\sigma = 0.2$	13.5%	38.8%	83.5%	42.1%	27.9%	70.5%	69.3%
stAdv, $\sigma = 0.5$	52.3%	84.4%	77.9%	81.7%	57.3%	67.3%	69.0%
stAdv, $\sigma = 2.0$	17.4%	30.6%	52.1%	27.0%	11.4%	27.2%	31.3%
l_p **robustness**							
$l_2, \epsilon = 0.5$	0.3%	39.2%	54.5%	58.3%	7.9%	43.8%	47.7%
$l_2, \epsilon = 1.0$	0.0%	9.5%	25.1%	29.7%	0.5%	16.8%	22.5%
$l_\infty, \epsilon = 1/255$	26.2%	84.4%	76.6%	83.5%	0.8%	20.1%	25.0%
$l_\infty, \epsilon = 4/255$	0.4%	39.8%	49.6%	68.3%	0.0%	0.1%	0.1%
$l_\infty, \epsilon = 8/255$	0.0%	10.3%	20.0%	45.4%	0.0%	0.0%	0.0%

Table 1: The performance of the models we considered under various noise distributions, together with our measurements of those models' robustness to small l_p perturbations. For all the robustness tests we used PGD with 100 steps and a step size of $\epsilon/25$. The adversarially trained CIFAR-10 model is the open sourced model from Madry et al. (2017). Unfortunately no open sourced robust ImageNet model was available to us at the time of writing.

Method	InceptionV3 (Undefended)	tv-min	JPEG	bitdepth	Random	HGD	Pixel
Noise Type							
Clean	76.0%	76.0%	74.3%	71.9%	72.6%	77.2%	73.1%
Gaussian, $\sigma = 0.1$	68.4%	71.5%	65.5%	67.5%	66.0%	68.1%	70.4%
Gaussian, $\sigma = 0.2$	52.0%	66.9%	50.0%	51.2%	52.0%	50.6%	53.2%

Table 2: The performance in Gaussian noise of several previously published defenses for ImageNet. All of these defenses are now known not to improve adversarial robustness (Athalye et al., 2018). The defense strategies include total variance minimization Guo et al. (2017), JPEG compression (Guo et al., 2017; Dziugaite et al., 2016; Liu et al., 2018; Aydemir et al., 2018; Das et al., 2018; 2017), bitdepth reduction (Guo et al., 2017), random resizing and random padding of the input image (Xie et al., 2017), respresentation guided denoising (Liao et al., 2018), and Pixel Deflection (Prakash et al., 2018).

6 Comparing Adversarial Training to Training on Noise

For a linear model, improving generalization in the presence of noise is equivalent to increasing the distance to the decision boundary. The results from the previous two sections suggest that a similar relationship should hold for other statistical classifiers, including neural networks. This suggests that augmenting the training data distribution with noisy images ought to increase the distance to the decision boundary, and augmenting the training distribution with small-perturbation adversarial examples should improve performance in noise. In this section we present evidence that it does.

We analyzed the performance of the models described in Section 3 on four different noise distributions: two types of Gaussian noise, pepper noise (Hendrycks & Dietterich, 2018), and a randomized variant of the stAdv adversarial attack introduced in Xiao et al. (2018). We used both ordinary, spherical Gaussian noise and what we call "PCA noise," which is Gaussian noise supported only on the subspace spanned by the first 100 principal components of the training set. Pepper noise randomly assigns channels of the image to 1 with some fixed probability. Details of the stAdv attack can be found in Appendix B, however it visually similar to Gaussian blurring where σ controls the severity of the blurring. Example images that have undergone each of the noise transformations we used can be found in Appendix H. Each model was also tested for l_p robustness with a variety of norms and ϵ's using the same PGD attack as in Section 3.

For CIFAR-10, standard Gaussian data augmentation yields comparable results to adversarial training on all considered metrics. For ImageNet we found that Gaussian data augmentation improves robustness to small l_2 perturbations as well as robustness to other noise corruptions. The results are shown in Table 1. This holds both for generalization in all noises considered and for robustness to small perturbations. We found that performing data augmentation with heavy Gaussian noise ($\sigma = 0.4$ for CIFAR-10 and $\sigma = 0.8$ for ImageNet) worked best. The adversarially trained CIFAR-10 models were trained in the l_∞ metric and they performed especially well on worst-case perturbations in this metric. Prior work has observed that Gaussian data augmentation helps small perturbation robustness on MNIST (Kannan et al., 2018), but to our knowledge we are the first to measure this on CIFAR-10 and ImageNet.

Neither augmentation method shows much improved generalization in PCA noise. We hypothesize that adversarially trained models learn to project away the high-frequency information in the input, which would do little to improve performance in PCA noise, which is supported in the low-frequency subspace of the data distribution. Further work would be required to establish this.

We also considered the MNIST adversarially trained model from Madry et al. (2017), and found it to be a special case where although robustness to small perturbations was increased generalization in noise was not improved. This is because this model violates the linearity assumption discussed in Section 5. This overfitting to the l_∞ metric has been observed in prior work (Sharma & Chen, 2017). More details can be found in Appendix D.

Although no l_p-robust open sourced model exists, recent work though has found that the adversarially trained models on Tiny ImageNet from Kannan et al. (2018) generalize very well on a large suite of common image corruptions (Hendrycks & Dietterich, 2018).

Failed Adversarial Defenses Do Not Improve Generalization in Noise. We performed a similar analysis on seven previously published adversarial defense strategies. These methods have already been shown to result in masking gradients, which cause standard optimization procedures to fail to find errors, rather than actually improving small perturbation robustness (Athalye et al., 2018). We find that these methods also show no improved generalization in Gaussian noise. The results are shown in Table 2. Given how easy it is for a method to show improved robustness to standard optimization procedures without changing the decision boundary in any meaningful way, we strongly recommend that future defense efforts evaluate on out of distribution inputs such as the noise distributions we consider here. The current standard practice of evaluating solely on gradient based attack algorithms is making scientific progress extremely difficult to measure.

Obtaining Zero Test Error in Noise is Nontrivial. It is important to note that applying Gaussian data augmentation does not reduce error rates in Gaussian noise to zero. For example, we performed Gaussian data augmentation on CIFAR-10 at $\sigma = .15$ and obtained 99.9% training accuracy but 77.5% test accuracy in the same noise distribution. (For comparison, the naturally trained obtains 95% clean test accuracy.) Previous work (Dodge & Karam, 2017b) has also observed that obtaining perfect generalization in large Gaussian noise is nontrivial. This mirrors Schmidt et al. (2018), which found that small perturbation robustness did not generalize to the test set. This is perhaps not surprising given that error rates on the *clean* test set are also non-zero. Although the model is in some sense "superhuman" with respect to clean test accuracy, it still makes many mistakes on the clean test set that a human would never make. We collected some examples in Appendix H. More detailed results on training and testing in noise can be found in Appendices C and G.

7 Conclusion

We proved a fundamental relationship between generalization in noisy image distributions and the existence of small adversarial perturbations. By appealing to the Gaussian isoperimetric inequality, we formalized the notion of what it means for a decision boundary to be badly behaved. We showed that, for noisy images, there is very little room to improve robustness without also decreasing the volume of the error set, and we provided evidence that small perturbations of clean images can also be explained in a similar way. These results show that small-perturbation adversarial robustness is closely related to generalization in the presence of noise and that future defense efforts can measure progress by measuring test error in different noise distributions.

Indeed, several such noise distributions have already been proposed, and other researchers have developed methods which improve generalization in these distributions (Hendrycks & Dietterich, 2018; Dodge & Karam, 2017b;a; Vasiljevic et al., 2016; Zheng et al., 2016). Our work suggests that adversarial defense and improving generalization in noise involve attacking the same set of errors in two different ways — the first community tries to remove the errors on the boundary of the error set while the second community tries to reduce the volume of the error set. The isoperimetric inequality connects these two perspectives, and suggests that improvements in adversarial robustness should result in improved generalization in noise and vice versa. Adversarial training on small perturbations on CIFAR-10 also improved generalization in noise, and training on noise improved robustness to small perturbations.

In the introduction we referred to a question from Szegedy et al. (2014) about why we find errors so close to our test points while the test error itself is so low. We can now suggest an answer: despite what our low-dimensional visual intuition may lead us to believe, these errors are not in fact unnaturally close given the error rates we observe in noise. There is a sense, then, in which we simply haven't reduced the test error enough to expect to have removed most nearby errors.

While we focused on the Gaussian distribution, similar conclusions can be made about other distributions. In general, in high dimensions, the ϵ-boundary measure of a typical set is large even when its volume is small, and this observation does not depend on anything specific about the Gaussian distribution. The Gaussian distribution is a special case in that we can easily prove that *all* sets will have large ϵ-boundary measure. Mahloujifar et al. (2018) proved a similar theorem for a larger class of distributions. For other data distributions not *every* set has large ϵ-boundary measure, but under some additional assumptions it still holds that *most* sets do. An investigation of this relationship on the MNIST distribution can be found in Gilmer et al. (2018b, Appendix G).

We believe it would be beneficial for the adversarial defense literature to start reporting generalization in noisy image distributions, such as the common corruption benchmark introduced in Hendrycks & Dietterich (2018), rather than the current practice of *only* reporting empirical estimates of adversarial robustness. There are several reasons for this recommendation.

1. Measuring test error in noise is significantly easier than measuring adversarial robustness — computing adversarial robustness perfectly requires solving an NP-hard problem for every point in the test set (Katz et al., 2017). Since Szegedy et al. (2014), hundreds of adversarial defense papers have been published. To our knowledge, only one (Madry et al., 2017) has reported robustness numbers which were confirmed by a third party. We believe the difficulty of measuring robustness under the usual definition has contributed to this unproductive situation.

2. Measuring test error in noise would also allow us to determine whether or not these methods improve robustness in a trivial way, such as how the robust MNIST model learned to threshold the input, or whether they have actually succeeded in improving generalization outside the natural data distribution.

3. All of the failed defense strategies we examined failed to improve generalization in noise. For this reason, we should be highly skeptical of defense strategies that only claim improved l_p-robustness but do not demonstrate improved generalization outside the natural data distribution.

4. Finally, if the goal is improving the security of our models in adversarial settings, errors in the presence of noise are often problematic in their own right. A model for which an attacker can find an error by sampling noise at random is not secure in most realistic threat models. In fact, there are settings in which these errors are worse from a security standpoint, since they are so much easier to find. Attacks using simple random sampling of noise have in fact already been shown to succeed on deployed commercial ML systems (Hosseini et al., 2017).

The interest in measuring l_p robustness arose from a sense of surprise that errors could be found so close to correctly classified points. But from the perspective described in this paper, the phenomenon is less surprising. Statistical classifiers make a large number of errors outside the data on which they were trained, and small adversarial perturbations are simply the nearest ones.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Ayse Elvan Aydemir, Alptekin Temizel, and Tugba Taskaya Temizel. The effects of jpeg and jpeg2000 compression on attacks using adversarial examples. *arXiv preprint arXiv:1803.10418*, 2018.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Christer Borell. The Brunn-Minkowski inequality in Gauss space. *Inventiones mathematicae*, 30(2): 207–216, 1975.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *arXiv preprint arXiv:1705.07263*, 2017a.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 39–57. IEEE, 2017b.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108. ACM, 2004.

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. *arXiv preprint arXiv:1802.06816*, 2018.

Samuel Dodge and Lina Karam. Quality resilient deep neural networks. *arXiv preprint arXiv:1703.08119*, 2017a.

Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *Computer Communication and Networks (ICCCN), 2017 26th International Conference on*, pp. 1–7. IEEE, 2017b.

Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, pp. 1632–1640, 2016.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. In *IEEE CVPR*, number CONF, 2018.

Jean-Yves Franceschi, Alhussein Fawzi, and Omar Fawzi. Robustness of classifiers to uniform ℓ_p and Gaussian noise. *arXiv preprint arXiv:1802.07971*, 2018.

Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018a.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018b.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018.

Hossein Hosseini, Baicen Xiao, and Radha Poovendran. Google's cloud vision api is not robust to noise. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*, pp. 101–105. IEEE, 2017.

Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pp. 97–117. Springer, 2017.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Jun Zhu, and Xiaolin Hu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

Zihao Liu, Qi Liu, Tao Liu, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. *arXiv preprint arXiv:1803.05787*, 2018.

Aleksander Madry, Aleksander Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial examples. *arXiv preprint arXiv:1706.06083*, 2017.

Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. *arXiv preprint arXiv:1809.03063*, 2018.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506–519. ACM, 2017.

Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8571–8580, 2018.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Mądry. Adversarially robust generalization requires more data. *arXiv preprint arXiv:1804.11285*, 2018.

Yash Sharma and Pin-Yu Chen. Breaking the madry defense model with l1-based adversarial examples. *arXiv preprint arXiv:1710.10733*, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL `http://arxiv.org/abs/1312.6199`.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. Examining the impact of blur on recognition by convolutional networks. *arXiv preprint arXiv:1611.05760*, 2016.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 4480–4488, 2016.

σ	0.00625	0.0125	0.025	0.075	0.15	0.25
Training Accuracy	100%	100%	100%	100%	99.9%	99.4%
Test Accuracy	96.0%	95.5%	94.8%	90.4%	77.5%	62.2%

Table 3: Wide ResNet-28-10 (Zagoruyko & Komodakis, 2016) trained and tested on CIFAR-10 with Gaussian noise with standard deviation σ.

σ	0	0.1	0.2	0.4	0.6	0.8
Clean Training Accuracy	91.5%	90.8%	89.9%	87.7%	86.1%	84.6%
Clean Test Accuracy	75.9%	75.5%	75.2%	74.2%	73.3%	72.4%
Noisy Training Accuracy	–	89.0%	85.7%	78.3%	71.7%	65.2%
Noisy Test Accuracy	–	73.9%	70.9%	65.2%	59.7%	54.0%

Table 4: The models from Section 3 trained and tested on ImageNet with Gaussian noise with standard deviation σ; the column labeled 0 refers to a model trained only on clean images.

A Training Details

Models trained on CIFAR-10. We trained the Wide-ResNet-28-10 model (Zagoruyko & Komodakis, 2016) using standard data augmentation of flips, horizontal shifts and crops in addition to Gaussian noise independently sampled for each image in every minibatch. The models were trained with the open-source code by Cubuk et al. (2018) for 200 epochs, using the same hyperparameters which we summarize here: a weight decay of 5e-4, learning rate of 0.1, batch size of 128. The learning rate was decayed by a factor of 0.2 at epochs 60, 120, 160.

Models trained on ImageNet. The ResNet-50 model (He et al., 2016) was trained with a learning rate of 1.6, batch size of 4096, and weight decay of 1e-4. During training, random crops and horizontal flips were used, in addition to the Gaussian noise independently sampled for each image in every minibatch. The models were trained for 90 epochs, where the learning rate was decayed by a factor of 0.1 at epochs 30, 60, and 80. Learning rate was linearly increased from 0 to the value of 1.6 over the first 5 epochs.

B Noise Attack Details

Here we provide more detail for the noise distributions considered in Section 6. The stAdv attack defines a flow field over the pixels of the image and shifts the pixels according to this flow. The field is parameterized by a latent Z. When we measure accuracy against our randomized variant of this attack, we randomly sample Z from a multivariate Gaussian distribution with standard deviation σ. To implement this attack we used the open sourced code from Xiao et al. (2018). PCA-100 noise first samples noise from a Gaussian distribution $\mathcal{N}(0, \sigma)$, and then projects this noise onto the first 100 PCA components of the data. For ImageNet, the input dimension is too large to perform a PCA decomposition on the entire dataset. So we first perform a PCA decomposition on 30x30x1 patches taken from different color channels of the data. To general the noise we first sample from a 900 dimensional Gaussian, then project this into the basis spanned by the top 100 PCA components, then finally tile this projects to the full 299x299 dimension of the input. Each color channel is constructed independently in this fashion.

C Training and Testing on Gaussian Noise

In Section 6, we mentioned that it is not trivial to learn the distribution of noisy images simply by augmenting the training data distribution. In Tables 3 and 4 we present more information about the performance of the models we trained and tested on various scales of Gaussian noise.

Model	Clean Accuracy	Pepper $p = 0.2$ Accuracy	Gaussian $\sigma = 0.3$ Accuracy	stAdv $\sigma = 1.0$ Accuracy	PCA-100 $\sigma = 0.3$ Accuracy
Clean	99.2%	81.4%	96.9%	89.5%	63.3%
Adv	98.4%	27.5%	78.2%	93.2%	47.1%

Table 5: The performance of ordinarily and adversarially trained MNIST models on various noise distributions.

D RESULTS ON MNIST

MNIST is a special case when it comes to the relationship between small adversarial perturbations and generalization in noise. Indeed prior has already observed that an MNIST model can trivially become robust to small l_∞ perturbations by learning to threshold the input (Schmidt et al., 2018), and observed that the model from Madry et al. (2017) indeed seems to do this. When we investigated this model in different noise distributions we found it generalizes worse than a naturally trained model, results are shown in Table 5. Given that it is possible for a defense to overfit to a particular l_p metric, future work would be strengthened by demonstrating improved generalization outside the natural data distribution.

E THE GAUSSIAN ISOPERIMETRIC INEQUALITY

A more common statement of the Gaussian isoperimetric inequality than the one we used in Section 4 is as follows:

Theorem. *Consider the standard normal distribution q on $\mathbb{R}^n$, and let E be a measurable subset of $\mathbb{R}^n$. Write*

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp(x^2/2) dx,$$

the cdf of the one-variable standard normal distribution.

For a measurable subset $E \subseteq \mathbb{R}^n$, write $\alpha(E) = \Phi^{-1}(\mathbb{P}_{x\sim q}[x \in E])$. Then for any $\epsilon \geq 0$,

$$\mathbb{P}_{x\sim q}[x \in E_\epsilon] \geq \Phi(\alpha(E) + \epsilon).$$

The version we stated in the text involved $\epsilon_q^*(E)$, the median distance from a random sample from q to the closest point in E. This is the same as the smallest ϵ for which $\mathbb{P}_{x\sim q}[x \in E_\epsilon] = \frac{1}{2}$. So, when $\epsilon = \epsilon_q^*(E)$, the left-hand side of the Gaussian isoperimetric inequality is $\frac{1}{2}$, giving us that $\Phi(\alpha + \epsilon_q^*(E)) \leq \frac{1}{2}$.

Since Φ^{-1} is a strictly increasing function, applying it to both sides preserves the direction of this inequality. But $\Phi^{-1}(\frac{1}{2}) = 0$, so we in fact have that $\epsilon_q^*(E) \leq -\alpha$, which is the statement we wanted.

F CHURCH WINDOW PLOTS

In this section we include many more visualizations of the sorts of church window plots we discussed briefly in Section 5. We will show an ordinarily trained model's predictions on several different slices through the same CIFAR test point which illustrate different aspects of the story told in this paper. These images are best viewed in color.

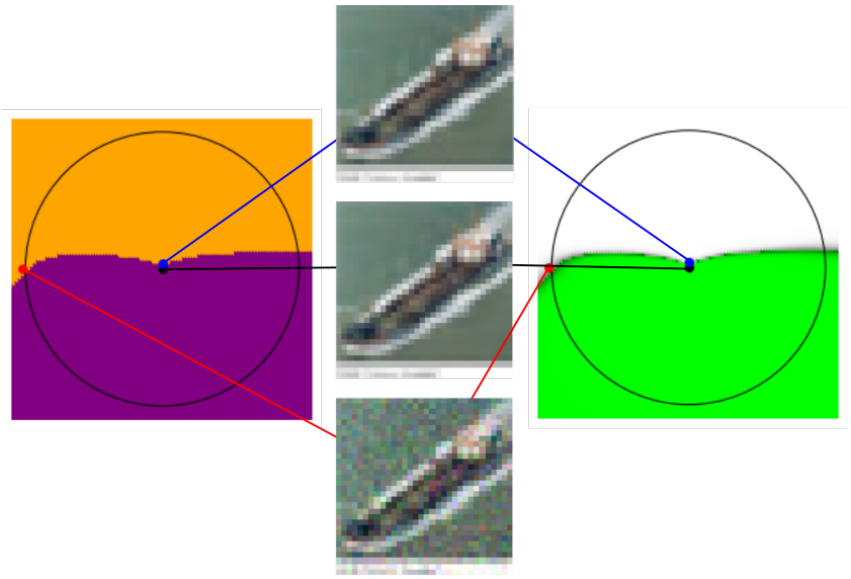

Figure 6: A slice through a clean test point (black, center image), the closest error found using PGD (blue, top image), and a random error found using Gaussian noise (red, bottom image). For this visualization, and all others in this section involving Gaussian noise, we used noise with $\sigma = 0.05$, at which the error rate was about 1.7%. In all of these images, the black circle indicates the distance at which the typical such Gaussian sample will lie. The plot on the right shows the probability that the model assigned to its chosen class. Green indicates a correct prediction, gray or white is an incorrect prediction, and brighter means more confident.

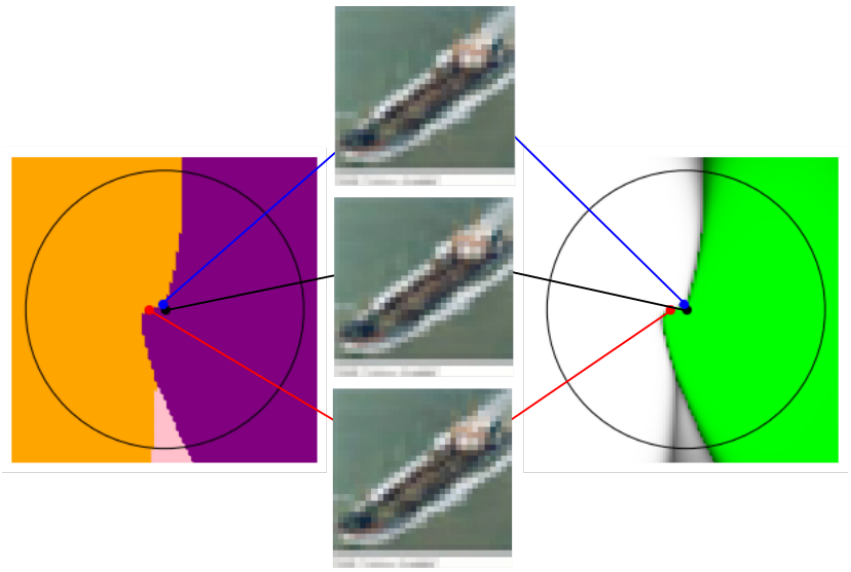

Figure 7: A slice through a clean test point (black, center image), the closest error found using PGD (blue, top image), and the average of a large number of errors randomly found using Gaussian noise (red, bottom image). The distance from the clean image to the PGD error was 0.12, and the distance from the clean image to the averaged error was 0.33. The clean image is assigned the correct class with probability 99.9995% and the average and PGD errors are assigned the incorrect class with probabilities 55.3% and 61.4% respectively. However, it is clear from this image that moving even a small amount into the orange region will increase these latter numbers significantly. For example, the probability assigned to the PGD error can be increased to 99% by moving it further from the clean image in the same direction by a distance of 0.07.

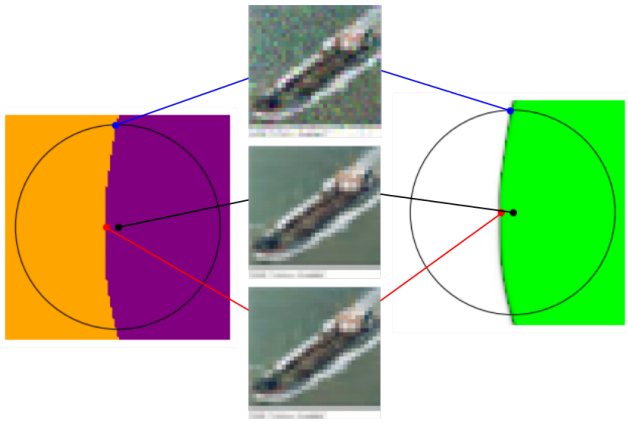

Figure 8: A slice through a clean test point (black, center image), a random error found using Gaussian noise (blue, top image), and the average of a large number of errors randomly found using Gaussian noise (red, bottom image).

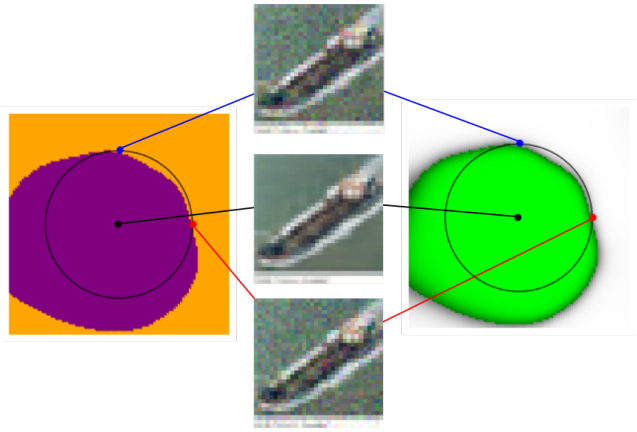

Figure 9: A slice through a clean test point (black, center image) and two random errors found using Gaussian noise (blue and red, top and bottom images). Note that both random errors lie very close to the decision boundary, and in this slice the decision boundary does not appear to come close to the clean image.

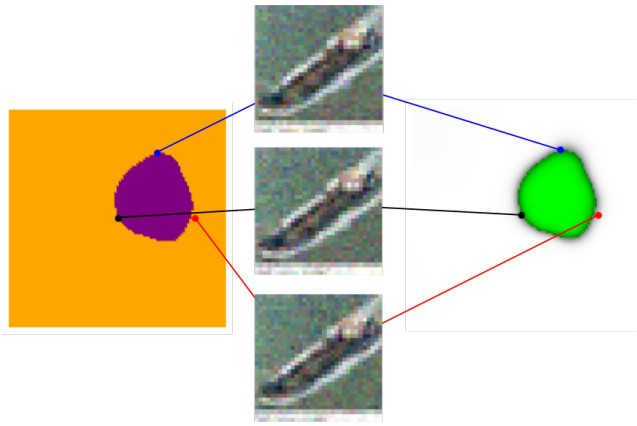

Figure 10: A slice through three random errors found using Gaussian noise. (Note, in particular, that the black point in this visualization does not correspond to the clean image.)

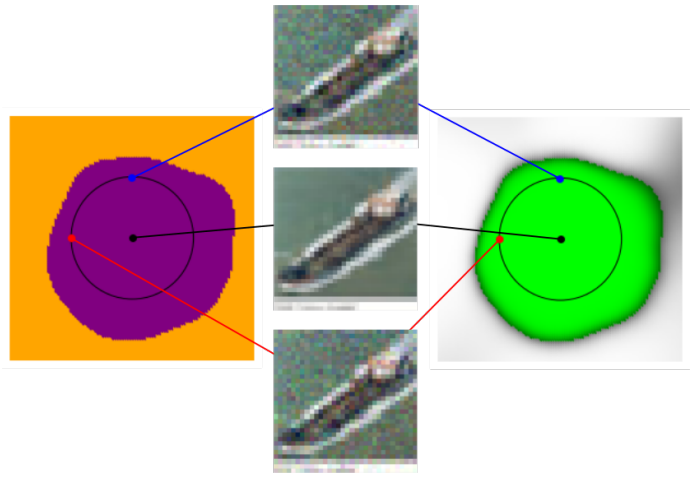

Figure 11: A completely random slice through the clean image.

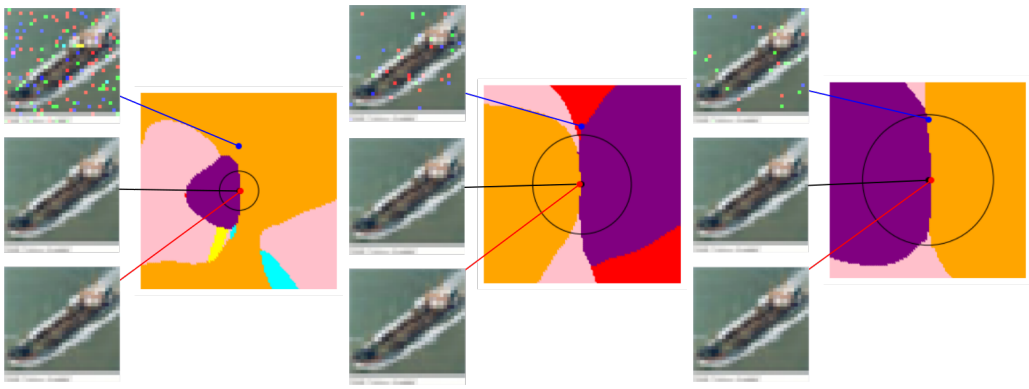

Figure 12: Some visualizations of the same phenomenon, but using the "pepper noise" discussed in Section 6 rather than Gaussian noise. In all of these visualizations, we see the slice through the clean image (black, center image), the same PGD error as above (red, bottom image), and a random error found using pepper noise (blue, top image). In the visualization on the left, we used an amount of noise that places the noisy image further from the clean image than in the Gaussian cases we considered above. In the visualization in the center, we selected a noisy image which was assigned to neither the correct class nor the class of the PGD error. In the visualization on the right, we selected a noisy image which was assigned to the same class as the PGD error.

G THE DISTRIBUTION OF ERROR RATES IN NOISE

Using some of the models that were trained on noise, we computed, for each image in the CIFAR test set, the probably that a random Gaussian perturbation will be misclassified. A histogram is shown in Figure 13. Note that, even though these models were trained on noise, there are still many errors around most images in the test set. While it would have been possible for the reduced performance in noise to be due to only a few test points, we see clearly that this is not the case.

H A COLLECTION OF MODEL ERRORS

In this section we first show a collection of iid test errors for the ResNet-50 model on the ImageNet validation set. We also visualize the severity of the different noise distributions considered in this work, along with model errors found by random sampling in these distributions.

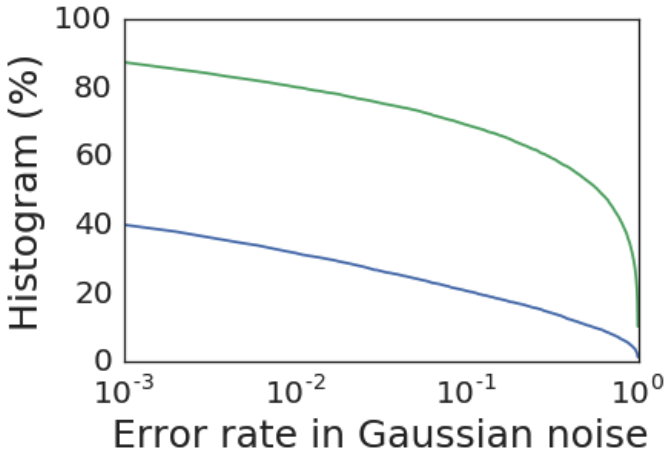

Figure 13: The cdf of the error rates in noise for images in the test set. The blue curve corresponds to a model trained and tested on noise with $\sigma = 0.1$, and the green curve is for a model trained and tested at $\sigma = 0.3$. For example, the left most point on the blue curve indicates that about 40% of test images had an error rate of at least 10^{-3}.

Figure 14: A collection of adversarially chosen model errors. These errors appeared in the ImageNet validation set. Despite the high accuracy of the model there remain plenty of errors in the test set that a human would not make.

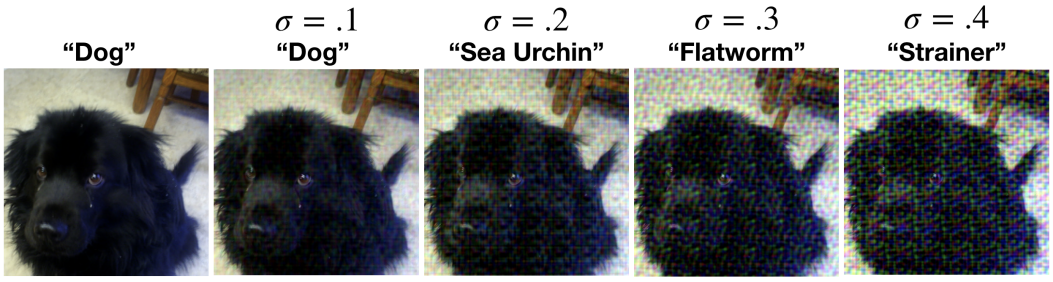

Figure 15: Visualizing the severity of PCA noise, along with model errors found in this noise distribution.

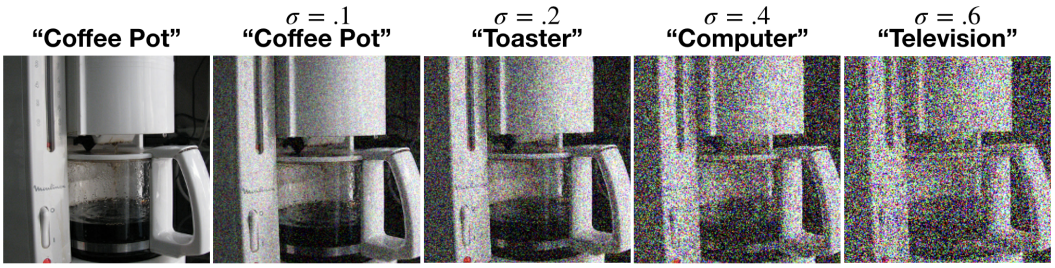

Figure 16: Visualizing the severity of Gaussian noise, along with model errors found in this noise distribution. Note the model shown here was trained at noise level $\sigma = .6$.

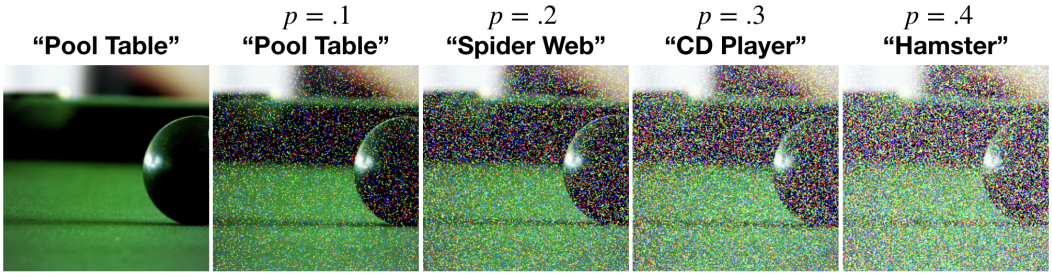

Figure 17: Visualizing the severity of pepper noise.

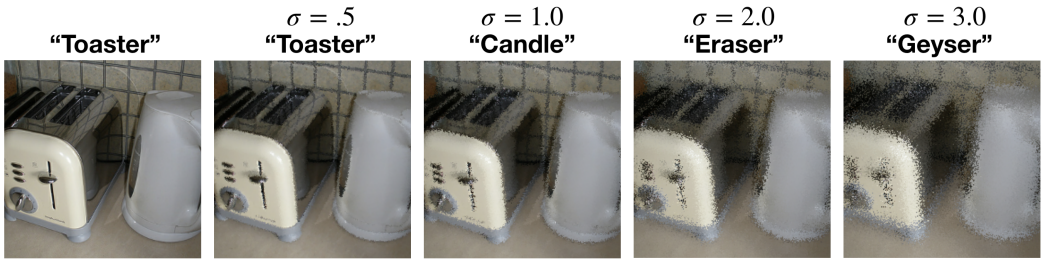

Figure 18: Visualizing the severity of the randomized stAdv attack.