Adversarial vulnerability for any classifier

Alhussein Fawzi[*] Hamza Fawzi[†] Omar Fawzi[‡]

February 23, 2018

Abstract

Despite achieving impressive and often superhuman performance on multiple benchmarks, state-of-the-art deep networks remain highly vulnerable to perturbations: adding small, imperceptible, adversarial perturbations can lead to very high error rates. Provided the data distribution is defined using a generative model mapping latent vectors to datapoints in the distribution, we prove that *no classifier* can be robust to adversarial perturbations when the latent space is sufficiently large and the generative model sufficiently smooth. Under the same conditions, we prove the existence of adversarial perturbations that transfer well across different models with small risk. We conclude the paper with experiments validating the theoretical bounds.

1 Introduction

Deep neural networks are powerful models that achieve state-of-the-art performance across several domains, such as bioinformatics [SEC15, CSB14], speech [HDY+12], and computer vision [HZRS15, KSH12]. Though deep networks have exhibited very good performance in classification tasks, they have recently been shown to be unstable to adversarial perturbations of the data [SZS+14, BCM+13]. In fact, very small and often imperceptible perturbations of the data samples are sufficient to fool state-of-the-art classifiers and result in incorrect classification. This discovery of the surprising vulnerability of classifiers to perturbations has led to a large body of work that attempts to design robust classifiers. However, advances in designing robust classifiers have been accompanied with stronger perturbation schemes that easily defeat such defenses [CW17, Bre17]. In fact, there is, to this date, no successful and scalable strategy to defend against adversarial perturbations. This leads to the following natural question:

Is it possible to design robust classifiers against adversarial perturbations?

Our main result is to prove that if the data distribution is defined by a smooth generative model with a sufficiently large latent space, then no classifier can be robust to adversarial noise. In addition, we prove the existence of perturbations that are transferable between classifiers having small risk, an intriguing property which was empirically observed in [SZS+14].

Specifically, we establish upper bounds on the classifiers' robustness to perturbations, which are applicable to *any* classification function. Our main assumption is the existence of a generative model that transforms normally-distributed random latent vectors into natural data. This assumption is motivated by numerous previous works on generative modelling, whereby natural-looking images

[*]DeepMind, London, UK.
[†]University of Cambridge, UK.
[‡]ENS Lyon, France.

are obtained by transforming normal vectors through a deep neural network [KW13], [GPAM+14], [RMC15], [ACB17], [GAA+17]. Under this data model, our classifier-agnostic bounds provide limits to the maximal robustness one can hope to achieve. Specifically, we first focus on the robustness to perturbations in the latent space; assuming the latent space is sufficiently high dimensional, we prove for any classifier the existence of small perturbations in the latent space that cause misclassification. Then, we show that the existence of such perturbations in the latent space implies the existence of small adversarial perturbations in the image space, provided the generator satisfies some smoothness condition (e.g., small Lipschitz constant). Specifically, this gives us fundamental upper bounds on the robustness of any classifier to perturbations in the image space, provided the data follows our model. Our main technical tool is the Gaussian isoperimetric inequality, which gives lower bounds on the *widenings* of any arbitrary measurable sets in Gaussian space. It should be noted that our bounds hold even when the generative model does not exactly model the data distribution, but rather provides a good *approximation*. That is, assuming that the data distribution is δ-close in Wasserstein sense to the one obtained with the generative model, our upper bounds still hold. This assumption is inline with recent advances in generative models, whereby the generator provides a good approximation (in the Wasserstein sense) to the true distribution, but does not exactly fit it [ACB17].

Finally, we prove one of the most intriguing properties about adversarial perturbations; their transferability [SZS+14, LCLS16]. That is, we show that arbitrary classifiers with small risk will have shared adversarial perturbations in the latent space. The existence of transferable adversarial perturbations across different models has important implications from a security perspective, as it opens the door to designing adversarial perturbations for models that are not known to the adversary.

1.1 Related works

The vulnerability of classifiers to perturbations has first been highlighted in [BCM+13], and studied in the context of deep neural networks in [SZS+14]. It was proven in [FFF15, FMDF16] that for certain families of classifiers, there exists adversarial perturbations that cause misclassification of magnitude $O(1/\sqrt{d})$, where d is the data dimension, provided the robustness to random noise is fixed (which is typically the case if e.g., the data is normalized). In addition, fundamental limits on the robustness of classifiers were derived in [FFF15] for some simple classification families (linear, quadratic); these risk dependent upper bounds apply for all classification functions in the family, and highlight the existence of a tradeoff between classification robustness and risk (i.e., the more accurate the classifier is, the smaller the upper bound). In this work, we instead derive bounds that hold for any classification function, and assume a model on the data distribution (that is, the distribution results from the transformation of normal vectors through a generator network g). More recently, lower bounds on the adversarial robustness of different classification models have been shown in [HA17, PRGS17]. In contrast to such works, we stress that our goal is instead to show *upper bounds* on the robustness that are classifier-agnostic. That is, while previous works have focused on proving conditions under which the classifier *is* robust, our goal is to show that all classifiers are *not* robust, provided conditions on the data are met. While the causes behind the high vulnerability of state-of-the-art networks to perturbations are still not completely understood, several hypotheses have been proposed by analyzing simple classifiers and toy tasks in [GSS15, TG16, GMF+18]. Simultaneously, a large number of techniques have recently been proposed to improve the robustness of classifiers to perturbations, such as adversarial training [GSS15], robust optimization [SYN15, MMS+17], regularization [CBG+17], distillation [PMW+15], stochastic networks [AFDM16], ... Unfortunately, such recent techniques have been shown to fail whenever a more complex attack strategy is used [CW17, Bre17]. More recently, algorithms which provide provably robust classifiers [SND17, RSL18] have been proposed and tested on small datasets (e.g., MNIST). For large scale, high dimensional datasets, it remains however unclear whether it is possible to design more robust classification methods. Our goal here is to show that there exists upper bounds on the robustness that no classifier can surpass. Finally, we note that the robustness of generative models have been empirically analyzed in [KFS17, TTV16]. Unlike these works, our goal is not to assess the robustness of generative models to perturbations; instead, we analyze the robustness of classifiers when the data is generated according to such generative models.

2 Problem setting

Let $f : \mathbb{R}^m \to [K]$ be an arbitrary K-class classifier operating in the space $\mathbb{R}^m$. We will often refer to the space $\mathcal{X} := \mathbb{R}^m$ as the *space of images*, where the dimensionality m refers to the number of pixels in an image $x \in \mathcal{X}$. We assume the existence of a generative model g that maps latent vectors $z \in \mathcal{Z} := \mathbb{R}^d$ to the space of images $\mathcal{X}$. Specifically, to generate an image according to the distribution of natural images μ, we generate a random vector $z \sim \nu$, where $\nu = \mathcal{N}(0, I_d)$, and we apply the map g (i.e, the image is $g(z) \in \mathbb{R}^m$). In other words, the distribution of images μ is the pushforward $g_*(\nu)$ of $\nu := \mathcal{N}(0, I_d)$ by the map g. This is a standard procedure used to generate images with state-of-the-art generative models.[1]

Let

$$C_i(f) = \{z \in \mathcal{Z} : (f \circ g)(z) = i\}$$

denote the set of latent vectors in $\mathcal{Z}$ that generate images belonging to class i (according to classifier f). Note that $\{C_i(f)\}_i$ is a partition in the the $\mathcal{Z}$ space (i.e., $\bigcup_{i=1}^{K} C_i(f) = \mathcal{Z}$ and $C_i(f) \cap C_j(f) = \emptyset$ for $i \neq j$). When the classifier is clear from the context, we simply write $C_i = C_i(f)$. Note that $\nu(C_i)$ is the measure of C_i in $\mathcal{Z}$; when the classes are equiprobable, we have $\nu(C_i) = 1/K$, where we recall that K is the number of classes.

We analyse in this paper several notions of robustness, which we define below:

- **In-distribution robustness:** For natural images, latent vectors provide a decomposition of images into meaningful factors of variation, such as features of objects in the image, illumination, etc... Perturbations of latent vectors in $\mathcal{Z}$ measure the amount of change one needs to apply to such meaningful latent features to cause data misclassification. We endow the space of latent vectors $\mathcal{Z}$ with the ℓ_2 norm, and measure the smallest required perturbation in the $\mathcal{Z}$-space that causes misclassification. Specifically, we define $r_{\mathcal{Z}}$ as follows

$$r_{\mathcal{Z}}(z) = \min_{r \in \mathcal{Z}} \|r\|_2 \text{ s.t. } (f \circ g)(z) \neq (f \circ g)(z + r).$$

  In the image space, we define the in-distribution robustness $r_{\mathcal{X}}(z)$ as follows:

$$r_{\mathcal{X}}(z) = \min_{r \in \mathcal{Z}} \|g(z + r) - g(z)\| \text{ s.t. } (f \circ g)(z) \neq (f \circ g)(z + r),$$

  where $\|\cdot\|$ denotes an arbitrary norm on $\mathcal{X}$. We refer to this robustness setting as *in-distribution*, as the perturbed image, $g(z + r)$, is constrained to lie in the image of g, and hence belongs to the support of the distribution μ.

[1] Instead of sampling from $\mathcal{N}(0, I_d)$ in $\mathcal{Z}$, some generative models sample from the uniform distribution in $[-1, 1]^d$. The results of this paper can be easily extended to such generative procedures.

- **Unconstrained robustness:** Unlike the in-distribution setting, we measure here the robustness to *arbitrary* perturbations in the image space; that is, the perturbed image is not constrained anymore to belong to the data distribution μ.

$$r^*(z) = \min_{r \in \mathcal{X}} \|r\| \text{ s.t. } f(g(z) + r) \neq f(g(z)).$$

  Note that the perturbation vector r is not constrained to belong to the image of g. It is easy to see that this robustness definition is smaller than the in-distribution robustness; i.e., $r^*(z) \leq r_{\mathcal{X}}(z)$ for all $z \in \mathcal{Z}$. This notion of robustness corresponds to the widely used definition of adversarial perturbations.

In the remainder of the paper, our goal is to analyse limits on the maximal achievable robustness to adversarial perturbations, when measuring the robustness in the latent space (through $r_{\mathcal{Z}}$) or in the image space (through $r_{\mathcal{X}}$ and r^*). To interpret such limits on the robustness, it is crucial to compare adversarial perturbations to the right quantity. In the image space, perturbations are often compared to the norm of the image [SZS+14, FFF15, MDFF16], and a perturbation is deemed sufficiently small whenever the robustness is much smaller than the norm of the image (i.e., at least one order of magnitude). To interpret the robustness in the latent space, we follow here a similar methodology, and compare $r_{\mathcal{Z}}$ to $\|z\|_2$; the classifier is not robust to perturbations in the latent space when $r_{\mathcal{Z}}(z) \ll \|z\|_2$ for most z. Assuming all coordinates of $\mathcal{Z}$ encode different factors of variation in the image, the condition $r_{\mathcal{Z}}(z) \ll \|z\|_2$ indicates that tiny changes to the latent factors of variation can cause data misclassification.[2] The next section bounds the robustness quantities for arbitrary classifiers.

3 Bounds on adversarial robustness

3.1 Robustness in the $\mathcal{Z}$-space

We start by stating a very general bound and then derive two special cases to make more explicit the dependence on the distribution on the classes and on the number of classes.

Theorem 1. *Let $f : \mathbb{R}^m \to \{1, \ldots, K\}$ be an arbitrary classification function defined on the image space. Then,*

$$\mathbb{P}_z \left(r_{\mathcal{Z}}(z) \leq \epsilon \right) \geq \sum_{i=1}^{K} \left(\Phi(a_{\neq i} + \epsilon) - \Phi(a_{\neq i}) \right) , \tag{1}$$

where Φ is the cdf of $\mathcal{N}(0,1)$, and $a_{\neq i} = \Phi^{-1}\left(\nu\left(\bigcup_{j \neq i} C_j(f)\right)\right)$.

In particular, if for all i, $\nu(C_i(f)) \leq \frac{1}{2}$ (the classes are not too unbalanced), we have

$$\mathbb{P}_z \left(r_{\mathcal{Z}}(z) \leq \epsilon \right) \geq 1 - \sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2} . \tag{2}$$

[2]The comparison between $r_{\mathcal{Z}}(z)$ and $\|z\|_2$ is only meaningful if the generative mode is non-degenerate; that is, all latent coordinates of z have an effect on the appearance of the image $g(z)$. In fact, by superficially increasing the dimension d of the latent space with dummy variables having no effect on $g(z)$, the norm $\|z\|_2$ increases with dimension, while $r_{\mathcal{Z}}(z)$ remains constant. We assume in the remainder of the paper that the generative model is non-degenerate; i.e., that all latent coordinates have an effect on the image.

To see the dependence on the number of classes more explicitly, consider the setting where the classes are equiprobable, i.e., $\nu(C_i(f)) = \frac{1}{K}$ *for all* i, $K \geq 5$, *then*

$$\mathbb{P}_z\left(r_{\mathcal{Z}}(z) \leq \epsilon\right) \geq 1 - \sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2} e^{-\epsilon\sqrt{\log\left(\frac{K^2}{4\pi\log(K)}\right)}}. \quad (3)$$

This theorem is a consequence of the Gaussian isoperimetric inequality first proved in [Bor75] and [ST78]. The detailed proof can be found in the Appendix A.2.

Remark 1. Interpretation. Theorem 1 shows the existence of adversarial perturbations in the latent space that have a *constant* norm (with respect to the dimension d of the latent space). The normalized robustness is therefore given by $\frac{\|r_{\mathcal{Z}}(z)\|_2}{\|z\|_2} = O(d^{-1/2})$ with high probability, since $\|z\|_2 \approx \sqrt{d}$ for $z \sim \mathcal{N}(0, I_d)$. Hence, this proves the existence of small adversarial perturbations in the latent space (compared to the typical norm of latent vectors), in the setting where d is sufficiently large. For example, for $d = 100$, we have $\frac{\|r_{\mathcal{Z}}(z)\|_2}{\|z\|_2} \leq 1/10$ with probability ≈ 0.7. See Fig. 2a for an illustration.

Remark 2. Dependence on K. Theorem 1 shows an *increasing* probability of fooling with the number of classes K. In other words, it is easier to find adversarial perturbations in the setting where the number of classes is large, than for a binary classification task.[3] This dependence confirms empirical results whereby the robustness is observed to decrease with the number of classes. The dependence on K captured in our bounds is in contrast to previous bounds that showed decreasing probability of fooling the classifier, for larger number of classes [FMDF16].

Remark 3. Classification-agnostic bound. Our bounds hold for *any* classification function f, and is not specific to a family of classifiers. This is unlike the work of [FFF15] that establishes bounds on the robustness for specific classes of functions (e.g., linear or quadratic classifiers).

Remark 4. Tightness of the bound. Suppose that the classifier f separates the $\mathcal{Z}$ space into $C_1 = \{z : z_1 \geq 0\}$ and $C_2 = \{z : z_1 < 0\}$. Then, it is easy to see that

$$\mathbb{P}(r_{\mathcal{Z}}(z)) \leq \epsilon) = \mathbb{P}(z \in C_1, z_1 < \epsilon) + \mathbb{P}(z \in C_2, z_1 \geq -\epsilon) = 2(\Phi(\epsilon) - \Phi(0)),$$

which precisely corresponds to Eq. (1). In this case, the bound in Eq. (1) is therefore an equality. More generally, the bound in Eq. (1) is an *equality* if the classifier is linear in the $\mathcal{Z}$-space. This suggests that classifiers are maximally robust when the induced classification boundaries in the $\mathcal{Z}$-space are linear. We stress on the fact that boundaries in the $\mathcal{Z}$-space can be very different from the boundaries in the image space. In particular, as g is in general non-linear, f might be a highly *non-linear function* of the input space, while $z \mapsto (f \circ g)(z)$ is a linear function in z.

Following Remark 4, we now explicitly show through a toy example that a classifier which is not linear in the $\mathcal{Z}$-space can be significantly less robust than a linear one.

Example 1 (Checkerboard class partitions)**.** *Assume that* C_1 *and* C_2 *are given by:*

- $C_1 = \{(x_1, \dots, x_d) : \sum_{i=1}^d \lfloor x_i \rfloor \mod 2 = 0\}$,
- $C_2 = \mathbb{R}^d - C_1$.

See Fig. 1a for an illustration. Then, we have

$$\mathbb{P}_z\left(r_{\mathcal{Z}}(z) \leq \epsilon\right) \geq 1 - (1 - \epsilon)^d. \quad (4)$$

[3] We assume here equiprobable classes.

See Appendix A.3 for details on the calculation. Fig 1b compares the general bound in Theorem 1 to Eq. (4). As can be seen, in the checkerboard partition example, the probability of fooling converges much quicker to 1 (wrt ϵ) than the general result in Theorem 1. Hence, a classifier that creates many disconnected classification regions can be much more vulnerable to perturbations than a linear classifier in the latent space.

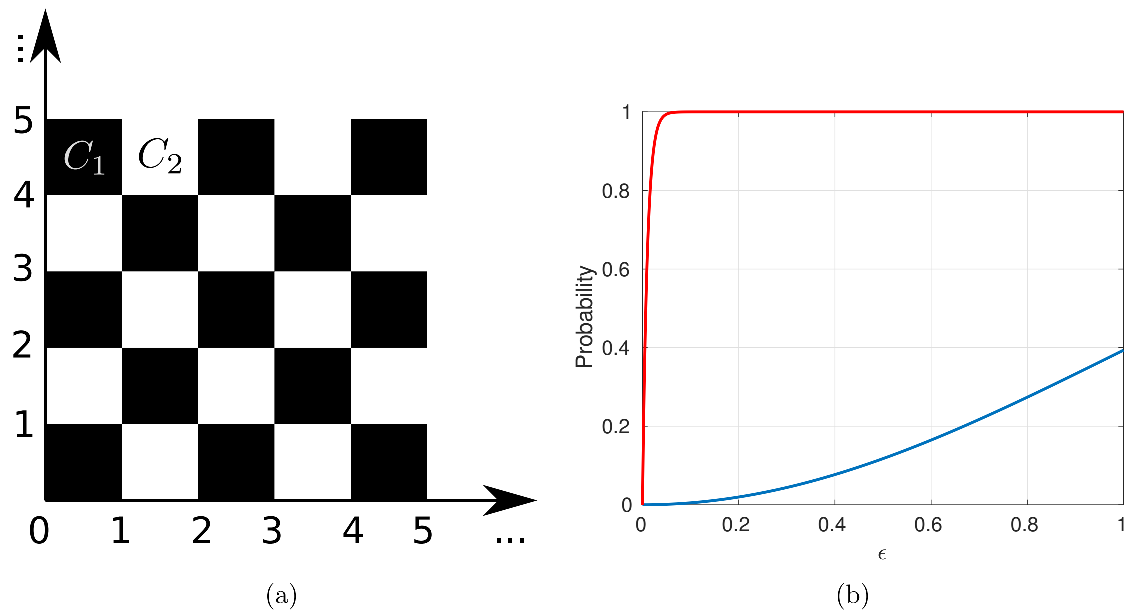


Figure 1: Left: Illustration of checkerboard example. Right: Lower bound on $\mathbb{P}(\|r_{\mathcal{Z}}(z)\|_2 \leq \epsilon)$ as a function of ϵ for the general result in Theorem 1 (blue curve) and the checkerboard example in Eq. 4 (red curve).

We now derive a bound on the expectation of the robustness to adversarial perturbations in the $\mathcal{Z}$ space.

Theorem 2. *With the same notations as in Theorem 1, we have:*

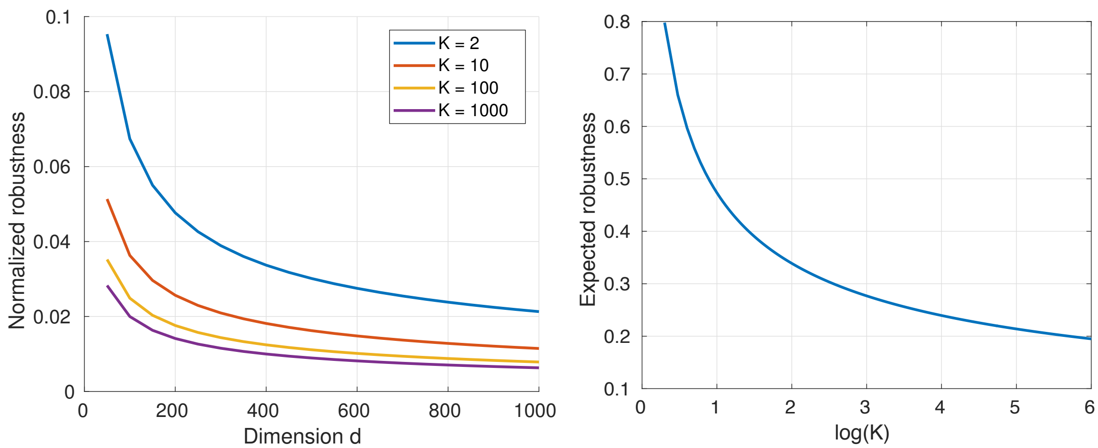
$$\mathbb{E}_z r_{\mathcal{Z}}(z) \leq \sum_{i=1}^{K} -a_{\neq i}\Phi(-a_{\neq i}) + \frac{e^{-a_{\neq i}^2/2}}{\sqrt{2\pi}}.$$

In particular, for $K \geq 5$ equiprobable classes, we have

$$\mathbb{E}_z r_{\mathcal{Z}}(z) \leq \frac{\log(4\pi \log(K))}{\sqrt{2\log(K)}}.$$

The upper bound on the expected robustness in the $\mathcal{Z}$-space should be compared to $\sqrt{d}$, the expected norm of a latent vector drawn from $\mathcal{N}(0, I_d)$. As a numerical example, assuming that the class measures are equal in the $\mathcal{Z}$ space (i.e., $\nu(C_i) = 1/K$), and $K = 10$, we therefore obtain $\mathbb{E}_z r_{\mathcal{Z}}(z) \leq 0.48 \ll \sqrt{d}$, for d taking a typical value (e.g., $d = 100$). Fig. 2b reports the numerical values for the upper bound, as a function of the number of classes K. Note that, similarly to our observation in Theorem 1, the upper bound gets smaller when increasing the number of classes.

We now turn to transferable adversarial perturbations. In the following, we show that two models with approximately zero risk will have shared adversarial perturbations. That is, there exist small perturbations that fool both classifiers with high probability.



(a) Upper bound (Theorem 1) on the median of the normalized robustness $r_{\mathcal{Z}}/\sqrt{d}$ for different values of K. We assume in this plot that classes have equal measure (i.e., $\nu(C_i) = 1/K$).

(b) Upper bound on the expected robustness, as a function of $\log_{10}(K)$, in the setting where the classes have equal measure (i.e., $\nu(C_i) = 1/K$). Note that the upper bound is decreasing, when increasing the number of classes K.

Figure 2

Theorem 3 (Transferability of perturbations)**.** *Let f, h be two classifiers. Assume that $\mathbb{P}_z(f \circ g(z) \neq h \circ g(z)) \leq \delta$ (e.g., if f and h have a risk bounded by $\delta/2$ for the data set generated by g). In addition, assume that $\nu(C_i(f)) + \delta \leq \frac{1}{2}$ for all i.*[4] *Then,*

$$\mathbb{P}_z \left(\exists v \in \mathcal{Z} : \|v\|_2 \leq \epsilon, f(g(z+v)) \neq f(g(z)), h(g(z+v)) \neq h(g(z)) \right) \geq 1 - \sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2} - 2\delta \ .$$

Compared to Theorem 1 which bounds the robustness to adversarial perturbations, the extra price to pay here to find *transferable* adversarial perturbations is the 2δ term, which is small if the risk of both classifiers is small. Hence, our bounds provide a theoretical explanation for the existence of transferable adversarial perturbations, which were previously shown to exist in [SZS+14, LCLS16]. The existence of transferable adversarial perturbations across several models with small risk has important security implications, as adversaries can, in principle, fool different classifiers with a single, classifier-agnostic, perturbation. The existence of such perturbations significantly reduces the difficulty of attacking (potentially black box) machine learning models.

3.2 Robustness in the image space

Our goal is now to derive bounds on the robustness in the image space. To do so, a fundamental ingredient is to relate distances in the latent space to distances in the image space. We do so by assuming that the generative model g admits a modulus of continuity property:

Assumption 1. *We assume that g admits an invertible modulus of continuity ω; i.e.*[5]*,*

$$\forall z, z', \|g(z) - g(z')\| \leq \omega(\|z - z'\|_2). \tag{5}$$

[4]This assumption is only to simplify the statement, a general statement can be easily derived in the same way.

[5]This assumption can be extended to random z. For ease of exposition however, we use the deterministic assumption.

Note that the above assumption is milder than assuming Lipschitz continuity. Lipschitz continuity assumption corresponds to choosing $\omega(t)$ to be a linear function of t, which may be difficult to satisfy in practice.

Under this assumption, we can leverage the results in Section 3.1, and provide upper bounds on the robustness in the image space. In particular, a variant of Theorem 1 holds for the in-distribution robustness in the image space, provided ϵ is replaced by $\omega^{-1}(\epsilon)$. For clarity, we state some of the main results of Section 3.1 as follows:

Theorem 4. *Let $f : \mathbb{R}^m \to \{1, \dots, K\}$ be an arbitrary classification function defined on the image space. Then,*

$$\mathbb{P}_z \left(r_{\mathcal{X}}(z) \leq \epsilon \right) \geq \sum_{i=1}^{K} (\Phi(a_{\neq i} + \omega^{-1}(\epsilon)) - \Phi(a_{\neq i})) \ , \tag{6}$$

where Φ is the cdf of $\mathcal{N}(0,1)$, and $a_{\neq i} = \Phi^{-1}\left(\nu\left(\bigcup_{j\neq i} C_j(f)\right)\right)$. Moreover, more explicit bounds (similarly to Theorem 1) can be derived in a straightforward way.

In the case where we assume Lipschitz continuity of the generator function g, $\omega^{-1}(\epsilon)$ should be replaced with ϵ/L. Our high probability bounds derived on the latent space are therefore also valid to bound the robustness in the image space, provided the generator function admits a modulus of continuity. In the following, we extend and generalize the upper bound on the expectation in the image space.

Theorem 5. *We use the same notations as in Theorem 4. Provided ω is a concave function, we have*

$$\mathbb{E}_z r_{\mathcal{X}}(z) \leq \omega\left(\sum_{i=1}^{K} -a_{\neq i}\Phi(-a_{\neq i}) + \frac{e^{-a_{\neq i}^2/2}}{\sqrt{2\pi}}\right).$$

Assuming now that the generator g only provides a δ approximation of the true distribution μ in the Wasserstein sense; that is, $W(g_*(\nu), \mu) \leq \delta$, the following inequality holds provided ω is concave

$$\mathop{\mathbb{E}}_{x\sim\mu} r^*(x) \leq \omega\left(\sum_{i=1}^{K} -a_{\neq i}\Phi(-a_{\neq i}) + \frac{e^{-a_{\neq i}^2/2}}{\sqrt{2\pi}}\right) + \delta,$$

where $r^*(x)$ is the unconstrained robustness in the image space.

Theorem 5 provides bounds similar to Theorem 2 on the expected robustness in the *image space*, provided the generator admits a modulus of continuity. Moreover, Theorem 5 shows similar upper bounds on the robustness, even when the generative model only *approximates* (in the Wasserstein sense) the true unknown distribution. This relaxed assumption is of practical relevance, as state-of-the-art generative models are known to provide good approximations to the true distribution in the Wasserstein sense [ACB17], even if they do not model exactly the distribution of natural images. Theorem 5 provides a formal justification showing that our results also hold even in such cases where the generative model does not perfectly model the data distribution. Finally, we note that our transferability result holds in a straightforward way in the image space, provided the modulus of continuity assumption holds.

While the previous bounds are specifically looking at the in-distribution robustness, in many cases, we are interested in achieving *unconstrained* robustness; that is, the perturbed image is not constrained to belong to the data distribution (or equivalently to the range of g). It is easy to see that all our bounds derived for the in-distribution robustness $r_{\mathcal{X}}(z)$ also hold for the unconstrained robustness $r^*(z)$ since it clearly holds that $r^*(z) \leq r_{\mathcal{X}}(z)$. One may wonder whether it is possible to get a better upper bound on $r^*(z)$ directly. We show here that this is not possible if we require our bound to hold for any general classifier. Specifically, we construct a family of classifiers for which $r^*(z) \geq \frac{1}{2} r_{\mathcal{X}}(z)$, which we now present:

For a given classifier f in the image space, define the classifier $\tilde{f}$ constructed in a nearest neighbour strategy:

$$\tilde{f}(x) = f(g(z^*)) \quad \text{with} \quad z^* = \arg\min_z \|g(z) - x\|. \tag{7}$$

Note that $\tilde{f}$ behaves exactly in the same way as f on the image of g (in particular, it has the same risk and in-distribution robustness). We show here that it has an unconstrained robustness that is at least half of the in-distribution robustness of f.

Theorem 6. *For the classifier $\tilde{f}$, we have $\mathbb{E}_z r^*(z) \geq \frac{1}{2} \mathbb{E}_z r_{\mathcal{X}}(z)$.*

Proof. Let $x = g(z) \in \mathcal{X}$ and $x' \in \mathcal{X}$. Let z^* be such that $\tilde{f}(x') = f(g(z^*))$. By definition of $\tilde{f}$, we have $\|x' - g(z^*)\| \leq \|x' - g(z)\|$. As such, using the triangle inequality, we get

$$\begin{aligned} \|g(z) - g(z^*)\| &\leq \|g(z) - x'\| + \|x' - g(z^*)\| \\ &\leq 2\|g(z) - x'\| \ . \end{aligned}$$

Taking the minimum over all x' such that $\tilde{f}(x) \neq \tilde{f}(x')$, we obtain

$$r_{\mathcal{X}}(z) \leq 2r^*(z).$$

□

The above result therefore shows that, as far as classifier-agnostic bounds are concerned, there is no fundamental difference between in-distribution and unconstrained robustness. Specifically, if a classifier has in-distribution robustness r, then we can construct a classifier with unconstrained robustness $r/2$. Hence, our tight upper bounds derived for the in-distribution robustness are also (up to a factor of $\frac{1}{2}$) tight for the unconstrained setting. In other words, universal limits derived for both notions of robustness are essentially the same.

Finally, note that the procedure to construct $\tilde{f}$ in Eq. (7) suggests a way to robustify any classifier against out-distribution perturbations. Such a nearest neighbour strategy is useful when the in-distribution robustness is much larger than the unconstrained robustness, and allows essentially the latter to match the former. This approach has been recently implemented by others in [PS18] and shown to be effective to improve robustness.

4 Experimental evaluation

We evaluate our bounds on the SVHN dataset [NWC+11]. The dataset contains color images of house numbers, and the task is to classify the digit at the center of the image. The dataset contains $73,257$ training images, and $26,032$ test images (we do not use the images in the 'extra' set). We train a DCGAN [RMC15] generative model on this dataset, with a latent vector dimension $d = 100$, and further consider several neural networks architectures for classification. For each classifier, the

	2-Layer LeNet	ResNet-18	ResNet-101	Upper bound
Accuracy	11%	4.8%	4.2 %	-
Robustness in the $\mathcal{Z}$-space	6.1×10^{-3}	6.1×10^{-3}	6.6×10^{-3}	16×10^{-3}
Robustness in image space (in-distribution)	3.3×10^{-2}	3.1×10^{-2}	3.1×10^{-2}	36×10^{-2}
Robustness in image space (out-distribution)	0.39×10^{-2}	1.1×10^{-2}	1.4×10^{-2}	36×10^{-2}

Table 1: Experiments on SVHN. We report the 25% percentile of the robustness at each cell, where probabilities are computed either empirically, or theoretically (for the upper bound). More precisely, we report t such that $\mathbb{P}_z\left(r(z) \leq t\right) \geq 0.25$, where r refers either to $r_{\mathcal{Z}}$, $r_{\mathcal{X}}$ or r^*.

empirical robustness is computed using the approach in [MDFF16], and compared to our theoretical limit in Theorem 1 and 4. Results are reported in Table 1.

Observe first that the upper bound on the robustness predicted by Theorem 1 is close to the empirical robustness computed in the $\mathcal{Z}$-space, for the different tested classifiers. This suggests that our bounds in the $\mathcal{Z}$-space provide tight estimates of the robustness of state-of-the-art deep networks. To compute the upper bound of the robustness in the image-space, we estimate empirically the smoothness of the generator, and apply our bound in Theorem 1.[6] In the image space, the theoretical prediction from our classifier-agnostic bounds is one order of magnitude larger than the empirical estimates. Note however that the bound is non-vacuous, as it predicts the norm of the required perturbation to be approximately 1/3 of the norm of images. This potentially leaves room for improving the robustness in the image space. It should however be noted that, unlike the robustness in the $\mathcal{Z}$ space, the bound in the image space is not expected to be tight empirically, due to the extra smoothness inequality, which might have degraded the tightness of the bound.

Further comparisons of the figures between in-distribution and unconstrained robustness in the image space interestingly shows that for the simple LeNet architecture, a large gap exists between these two quantities. However, by using more complex classifiers (ResNet-18 and ResNet-101), the gap between in-distribution and unconstrained robustness gets smaller. Recall that Theorem 6 says that any classifier can be modified in a way that the in-distribution robustness and unconstrained robustness only differ by a factor 2, while preserving the accuracy. But this modification may be significantly more complicated than the original one, for example starting with a linear classifier, the modified classifier will in general not be linear. It is thus natural to expect that as we consider a more flexible family of classifiers, the gap between in-distribution and unconstrained robustness gets smaller.

5 Discussion

We showed the existence of classifier-agnostic upper bounds on the robustness to adversarial perturbations, assuming the data is generated according to a smooth generative model with sufficiently large latent space. Exceeding such bounds is not possible for *any* classifier; our goal should therefore be to approach such bounds. We list several implications of this paper:

1. In order to hope reaching this bound, we need to construct classifiers that act a linear classifier in the $\mathcal{Z}$ space. In particular, classifiers with multiple disconnected classification regions will be more prone to small perturbations (see Fig. 1b for an illustration).

[6] We used a probabilistic version of the modulus of continuity, where the property is not required to be satisfied for *all* z, z', but rather with high probability, and accounted for the error probability in the bound. We further used a gradient descent strategy to find the worst-case z' leading to largest $\|g(z) - g(z')\|$.

2. It is possible to "robustify" any classifier f provided we have access to a generative model by implementing the nearest neighbour strategy in Eq. (7). Such a classifier will (provably) close the gap between in-distribution and out-distribution robustness and hence achieve better robustness.

3. The unconstrained robustness upper bounds obtained from an empirical evaluation of our results are currently large (although clearly non-vacuous); i.e., one order of magnitude larger than the unconstrained robustness of state-of-the-art classifiers. This is possibly due to the low dimensionality of current generative models, which use $d \approx 100$, and working on simple datasets such as SVHN. Moving to more complex datasets will likely require larger latent space dimensions, and will hence lead to smaller bounds on the robustness of *all* classifiers to perturbations.

4. Our bounds are intriguingly applicable to *any* classification function, including the human visual system. It is however unclear how to interpret this result, and one should be cautious before concluding that humans are not robust to adversarial perturbations. We list here some (non-exhaustive) alternative explanations:

   - Natural distributions might not be possible to model as $g_*(\mathcal{N}(0, I_d))$, with sufficiently large d.
   - Humans might be achieving exactly this bound, which might be large for natural images.
   - The bounds are only on norms of the perturbations, and do not quantify the *perceptibility* of such perturbations. A possibility is that the norm of the perturbation is small, but that the perturbation is perceptible.

   Investigation of these hypotheses will be the subject of future works.

Acknowledgements

We would like to thank Seyed Mohsen Moosavi-Dezfooli, Neil Rabinowitz and Bernardino Romera-Paredes for useful feedbacks and discussions.

A Proofs

A.1 Useful results

Recall that we write the cumulative distribution function for the standard Gaussian distribution $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$. We state the Gaussian isoperimetric inequality [Bor75, ST78], the main technical tool used in to prove the results in this paper.

Theorem 7 (Gaussian isoperimetric inequality)**.** *Let γ_d be the Gaussian measure on $\mathbb{R}^d$. Let $A \subseteq \mathbb{R}^d$ and let $A_\eta = \{z \in \mathbb{R}^k : \exists z' \in A \text{ s.t. } \|z - z'\| \leq \eta\}$. If $\gamma_d(A) = \Phi(a)$ then $\gamma_k(A_\eta) \geq \Phi(a + \eta)$.*

We then state some useful bounds on the cumulative distribution function for the Gaussian distribution Φ.

Lemma 1 (see e.g., [Due10])**.** *We have for $x \geq 0$,*

$$1 - \frac{e^{-x^2/2}}{\sqrt{2\pi}} \frac{2}{x + \sqrt{x^2 + 8/\pi}} \leq \Phi(x) \leq 1 - \frac{e^{-x^2/2}}{\sqrt{2\pi}} \frac{2}{x + \sqrt{x^2 + 4}} \ . \tag{8}$$

Lemma 2. *Let $p \in [1/2, 1]$, we have for all $\epsilon > 0$,*

$$\Phi(\Phi^{-1}(p) + \epsilon) \geq 1 - (1-p)\sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2} e^{-\epsilon \Phi^{-1}(p)} \ . \tag{9}$$

If $p = 1 - \frac{1}{K}$ for $K \geq 5$ and $\epsilon \geq 1$, we have

$$\Phi(\Phi^{-1}(1 - \frac{1}{K}) + \epsilon) \geq 1 - \frac{1}{K}\sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2} e^{-\epsilon\sqrt{\log\left(\frac{K^2}{4\pi \log(K)}\right)}} \ . \tag{10}$$

Proof. As $p \geq 1/2$, we have $\Phi^{-1}(p) \geq 0$. Thus,

$$\begin{aligned}
\Phi(\Phi^{-1}(p) + \epsilon) &\geq 1 - \frac{1}{\sqrt{2\pi}} \frac{2e^{-(\Phi^{-1}(p)+\epsilon)^2/2}}{\Phi^{-1}(p) + \epsilon + \sqrt{(\Phi^{-1}(p)+\epsilon)^2 + 8/\pi}} \\
&= 1 - \frac{1}{\sqrt{2\pi}} \frac{2e^{-\Phi^{-1}(p)^2/2 - \epsilon^2/2 - \epsilon\Phi^{-1}(p)}}{\Phi^{-1}(p) + \epsilon + \sqrt{(\Phi^{-1}(p)+\epsilon)^2 + 8/\pi}} \\
&= 1 - \left(\frac{1}{\sqrt{2\pi}} \frac{2e^{-\Phi^{-1}(p)^2/2}}{\Phi^{-1}(p) + \sqrt{\Phi^{-1}(p)^2 + 4}}\right) \frac{\Phi^{-1}(p) + \sqrt{\Phi^{-1}(p)^2 + 4}}{\Phi^{-1}(p) + \epsilon + \sqrt{(\Phi^{-1}(p)+\epsilon)^2 + 8/\pi}} e^{-\epsilon^2/2 - \epsilon\Phi^{-1}(p)} \ .
\end{aligned}$$

Now we use the fact that

$$\left(\frac{1}{\sqrt{2\pi}} \frac{2e^{-\Phi^{-1}(p)^2/2}}{\Phi^{-1}(p) + \sqrt{\Phi^{-1}(p)^2 + 4}}\right) \leq 1 - \Phi(\Phi^{-1}(p)) = 1 - p \ .$$

As a result,

$$\begin{aligned}
\Phi(\Phi^{-1}(p) + \epsilon) &\geq 1 - (1-p)e^{-\epsilon^2/2 - \epsilon\Phi^{-1}(p)} \frac{\Phi^{-1}(p) + \sqrt{\Phi^{-1}(p)^2 + 4}}{\Phi^{-1}(p) + \epsilon + \sqrt{(\Phi^{-1}(p)+\epsilon)^2 + 8/\pi}} \\
&\geq 1 - (1-p)e^{-\epsilon^2/2 - \epsilon\Phi^{-1}(p)} \frac{\Phi^{-1}(p) + \sqrt{\Phi^{-1}(p)^2 + 4}}{\Phi^{-1}(p) + \sqrt{\Phi^{-1}(p)^2 + 8/\pi}} \\
&\geq 1 - (1-p)e^{-\epsilon^2/2} e^{-\epsilon\Phi^{-1}(p)} \frac{\sqrt{4}}{\sqrt{8/\pi}} \ .
\end{aligned}$$

In the case $p = 1 - \frac{1}{K}$, it suffices to show that that for $K \geq 5$, we have

$$\Phi^{-1}(1 - 1/K) \geq \sqrt{\log\left(\frac{K^2}{4\pi \log(K)}\right)} \ . \tag{11}$$

Using the upper bound in (8), it suffices to show that $\frac{1}{2}\frac{e^{-x^2}}{\sqrt{\pi}\sqrt{x^2+1}} \geq \frac{1}{K}$ where $x = \sqrt{\frac{1}{2}\log\left(\frac{K^2}{4\pi\log(K)}\right)}$. This inequality is equivalent to showing that $\sqrt{\log(K)} \geq \sqrt{x^2+1}$ for the same value of x. If we let $u = \log(K)$ this amounts to showing that $\sqrt{u} \geq \sqrt{u - \frac{1}{2}\log(4\pi u) + 1}$ for all $u \geq \log(5)$. For such u one can verify that $-\frac{1}{2}\log(4\pi u) + 1 \leq 0$ and so clearly the inequality is satisfied.

□

A.2 Proof of Theorem 1

Proof. To prove the general bound in Eq. (1), we define

$$C_{i\to}(\epsilon) = \{z \in C_i : \mathrm{dist}(z, \cup_{j\neq i} C_j) \leq \epsilon\}.$$

Here, $\mathrm{dist}(x, C)$ is defined as $\inf_{z'\in C} \|z - z'\|_2$. Note that for $z \in C_i$, $\|r_{\mathcal{Z}}(z)\|_2 \leq \epsilon$ if and only if $z \in C_{i\to}(\epsilon)$. Moreover, $C_{i\to}(\epsilon) \bigcup \cup_{j\neq i} C_j$ is nothing but the set of points that are at distance at most ϵ from $\cup_{j\neq i} C_j$. As such, by the Gaussian isoperimetric inequality (Theorem 7) applied with $A = \cup_{j\neq i} C_j$ and $a = a_{\neq i}$, we have $\nu(C_{i\to}(\epsilon)) + \nu(\cup_{j\neq i} C_j) \geq \Phi(a_{\neq i} + \epsilon)$, i.e., $\nu(C_{i\to}(\epsilon)) \geq \Phi(a_{\neq i} + \epsilon) - \Phi(a_{\neq i})$. As $C_{i\to}(\epsilon)$ are disjoint for different i, we have

$$\nu(\cup_i C_{i\to}(\epsilon)) \geq \sum_{i=1}^{K} (\Phi(a_{\neq i} + \epsilon) - \Phi(a_{\neq i}))\ .$$

This proves inequality (1).

To prove inequality (2), observe that if $\nu(C_i) \leq \frac{1}{2}$ for all i, then $\nu(\cup_{j\neq i} C_j) \geq \frac{1}{2}$ for all i. Then we use the bound (9) to get,

$$\begin{aligned}
\nu(\cup_i C_{i\to}(\epsilon)) &\geq \sum_{i=1}^{K} (\Phi(\Phi^{-1}(\nu(\cup_{j\neq i} C_j)) + \epsilon) - \nu(\cup_{j\neq i} C_j)) \\
&\geq \sum_{i=1}^{K} (1 - (1 - \nu(\cup_{j\neq i} C_j))\sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2} - \nu(\cup_{j\neq i} C_j)) \\
&= (1 - \sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2}) \sum_{i=1}^{K} (1 - \nu(\cup_{j\neq i} C_j)) \\
&= 1 - \sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2}\ .
\end{aligned}$$

For the bound (3) that makes explicit the dependence on the number of classes, we simply use the more explicit bound in (10). ☐

A.3 Proof of Example 1

Proof. We have $\nu(C_1) = \nu(C_2) = \frac{1}{2}$. Let $y \in \mathbb{R}^d$ in C_2 be such that for some $i \in \{1, \dots, d\}$, $y_i - \lfloor y_i \rfloor \in [0, \epsilon) \cup (1 - \epsilon, 1)$, then $y - \epsilon e_i \in C_1$ or $y + \epsilon e_i \in C_1$, and thus y is at distance at most ϵ from C_1. As a result, if y is at distance $> \epsilon$ from C_1, then for all $i \in \{1, \dots, d\}$, $y_i - \lfloor y_i \rfloor \in [\epsilon, 1 - \epsilon]$. As a result,

$$\begin{aligned}
\mathbb{P}_y(y \in C_2, \mathrm{dist}(y, C_1) > \epsilon) &\leq \mathbb{P}_y(y \in C_2, \forall i, y_i - \lfloor y_i \rfloor \in [\epsilon, 1 - \epsilon]) \\
&= \frac{1}{\sqrt{2\pi}^d} \sum_{(j_1,\dots,j_d)\in\mathbb{Z}^d, j_1+\cdots+j_d \mod 2 = 1} \int_{j_1+\epsilon}^{j_1+1-\epsilon} \cdots \int_{j_d+\epsilon}^{j_d+1-\epsilon} e^{-\frac{y_1^2+\cdots+y_d^2}{2}} dy_1 \dots dy_d\ .
\end{aligned}$$

Now observe that for any $j \in \mathbb{Z}$, as the function $y \mapsto e^{-y^2/2}$ is monotone on the interval $[j, j + 1]$ (nondecreasing if $j < 0$ and nonincreasing if $j \geq 0$). Thus, we have $\int_{j+\epsilon}^{j+1-\epsilon} e^{-\frac{y^2}{2}} dy \leq (1 -$

$\epsilon) \int_j^{j+1} e^{-\frac{y^2}{2}} dy$, when $\epsilon \leq \frac{1}{2}$. As a result,

$$\begin{aligned}
\mathbb{P}_y(y \in C_2, \mathrm{dist}(y, C_1) > \epsilon) &\leq \frac{1}{\sqrt{2\pi}^d}(1-\epsilon)^d \sum_{(j_1,\ldots,j_d)\in\mathbb{Z}^d, j_1+\cdots+j_d \mod 2=1} \int_{j_1}^{j_1+1} \cdots \int_{j_d}^{j_d+1} e^{-\frac{y_1^2+\cdots+y_d^2}{2}} dy_1 \ldots dy_d \\
&= (1-\epsilon)^d \mathbb{P}_y(y \in C_2) \\
&= \frac{1}{2}(1-\epsilon)^d \ .
\end{aligned}$$

With the same reasoning, $\mathbb{P}_y(y \in C_1, \mathrm{dist}(y, C_2) > \epsilon) \leq \frac{1}{2}(1-\epsilon)^d$ and gives inequality (4). $\square$

A.4 Proof of Theorem 3

Proof. We use the notation $\overline{C}$ for the complement of the set C. Define

$$C_{i\to}(\epsilon) = \{x \in C_i(f) \cup C_i(h) : \mathrm{dist}(x, \overline{C}_i(f) \cap \overline{C}_i(h)) \leq \epsilon\}.$$

Note that $C_i(f) \cup C_i(h) = \overline{\overline{C}_i(f) \cap \overline{C}_i(h)}$. We have $\nu(\overline{C}_i(f) \cap \overline{C}_i(h)) \geq \nu(\overline{C}_i(f)) - \delta = 1 - \nu(C_i(f)) - \delta \geq \frac{1}{2}$. Thus, using the Gaussian isoperimetric inequality with $A = \overline{C}_i(f) \cap \overline{C}_i(h)$, we obtain

$$\nu(C_{i\to}(\epsilon)) + \nu(\overline{C}_i(f) \cap \overline{C}_i(h)) \geq 1 - \left(1 - \nu(\overline{C}_i(f) \cap \overline{C}_i(h))\right) \sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2} \ ,$$

where we also used inequality (9). As a result,

$$\begin{aligned}
\nu(C_{i\to}(\epsilon)) &\geq (1 - \nu(\overline{C}_i(f) \cap \overline{C}_i(h))(1 - \sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2}) \\
&\geq \nu(C_i(f))(1 - \sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2}) \ .
\end{aligned}$$

Now assume that $z \in C_{i\to}(\epsilon)$ but also $z \in C_i(f) \cap C_i(h)$. Then it is classified as i for both f and h. In addition, the condition $z \in C_{i\to}(\epsilon)$ ensures that there exists $z' \in \overline{C}_i(f) \cap \overline{C}_i(h)$ such that $\|z - z'\|_2 \leq \epsilon$. Setting $v = z' - z$, we have that $f(g(z+v)) \neq f(g(z))$ and $h(g(z+v)) \neq h(g(z))$. As such it suffices to show that the set $C_{i\to}(\epsilon) \cap (C_i(f) \cap C_i(h))$ has sufficiently large measure. Indeed, we have

$$\nu(C_{i\to}(\epsilon) \cap (C_i(f) \cap C_i(h))) \geq \nu(C_{i\to}(\epsilon)) - \nu(C_i(f) \cap \overline{C}_i(h)) - \nu(\overline{C}_i(f) \cap C_i(h)) \ .$$

Summing over i, we get

$$\sum_{i=1}^{K} \nu(C_{i\to}(\epsilon) \cap (C_i(f) \cap C_i(h))) \geq 1 - \sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2} - 2\delta \ ,$$

because $\sum_{i=1}^{K} \nu(C_i(f) \cap \overline{C}_i(h)) + \nu(\overline{C}_i(f) \cap C_i(h)) = 2 \cdot \mathbb{P}_z(f \circ g(z) \neq h \circ g(z)) \leq 2\delta$.

$\square$

A.5 Proof of Theorem 5

Proof. For the first statement, we note that

$$\mathbb{E}_z r_{\mathcal{X}}(z) \leq \mathbb{E}_z \|g(z + \vec{r}_{\mathcal{Z}}(z)) - g(z)\| \leq \mathbb{E}_z \omega(r_{\mathcal{Z}}) \leq \omega\left(\mathbb{E} r_{\mathcal{Z}}(z)\right),$$

where we used the concavity of ω along with Jensen's inequality and $\vec{r}_{\mathcal{Z}}(z)$ denotes a perturbation of norm $r_{\mathcal{Z}}(z)$ that causes misclassification. The statement then follows using Theorem 2.

We assume now that g is such that $W(g_*(\nu), \mu) \leq \delta$, where W denotes the Wasserstein distance in $(\mathcal{X}, \|\cdot\|)$. Let (X, X') be a coupling with $X \sim \mu$ and $X' \sim g_*(\nu)$. We will construct a random variable X'' such that almost surely X'' and X are classified differently. We define $X'' = X'$ if X and X' are classified differently and otherwise $X'' = X' + \vec{r}^*(X')$ where $\vec{r}^*(X')$ is defined to be a vector of minimum norm such that $X' + \vec{r}^*(X')$ and X' are classified differently. Then we have

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mu} r^*(x) \leq \mathbb{E}\|X - X''\| &= \mathbb{E}(\mathbf{1}_{f(X)\neq f(X')}\|X - X'\|) + \mathbb{E}(\mathbf{1}_{f(X)=f(X')}\|X - (X' + \vec{r}^*(X'))\|) \\ &\leq \mathbb{E}\|X - X'\| + \mathbb{E}\|\vec{r}^*(X'))\| \ .\end{aligned}$$

By choosing a coupling such that $W(g_*(\nu), \nu) = \mathbb{E}\|X - X'\|$, we get $\mathbb{E}\|X - X'\| \leq \delta$. In addition, $\mathbb{E}\|\vec{r}^*(X'))\| \leq \mathbb{E}_{z\sim\nu} r_{\mathcal{X}}(z)$. The statement therefore follows. □

References

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. 1, 2, 8

[AFDM16] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 2

[BCM+13] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 387–402, 2013. 1, 2

[Bor75] Christer Borell. The Brunn-Minkowski inequality in Gauss space. *Inventiones mathematicae*, 30(2):207–216, 1975. 4, 10

[Bre17] Wieland Brendel. The robust vision benchmark. `http://robust.vision`, 2017. 1, 2

[CBG+17] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pages 854–863, 2017. 2

[CSB14] Davide Chicco, Peter Sadowski, and Pierre Baldi. Deep autoencoder neural networks for gene ontology annotation predictions. In *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 533–540, 2014. 1

[CW17] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017. 1, 2

[Due10] Lutz Duembgen. Bounding standard gaussian tail probabilities. *arXiv preprint arXiv:1012.2063*, 2010. 11

[FFF15] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations. *CoRR*, abs/1502.02590, 2015. 2, 4, 5

[FMDF16] Alhussein Fawzi, Seyed Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Neural Information Processing Systems (NIPS)*, 2016. 2, 5

[GAA+17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017. 1

[GMF+18] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018. 2

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[GSS15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 2

[HA17] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, pages 2263–2273, 2017. 2

[HDY+12] Geoffrey E. Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 29(6):82–97, 2012. 1

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 1

[KFS17] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. *arXiv preprint arXiv:1702.06832*, 2017. 2

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, pages 1097–1105, 2012. 1

[KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[LCLS16] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016. 2, 6

[MDFF16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 9

[MMS+17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2

[NWC+11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 5, 2011. 9

[PMW+15] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *arXiv preprint arXiv:1511.04508*, 2015. 2

[PRGS17] Jonathan Peck, Joris Roels, Bart Goossens, and Yvan Saeys. Lower bounds on the robustness to adversarial perturbations. In *Advances in Neural Information Processing Systems*, pages 804–813, 2017. 2

[PS18] Rama Chellappa Pouya Samangouei, Maya Kabkab. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018. 9

[RMC15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1, 9

[RSL18] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018. 2

[SEC15] Matt Spencer, Jesse Eickholt, and Jianlin Cheng. A deep learning network approach to ab initio protein secondary structure prediction. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 12(1):103–112, 2015. 1

[SND17] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017. 2

[ST78] Vladimir N Sudakov and Boris S Tsirel'son. Extremal properties of half-spaces for spherically invariant measures. *Journal of Soviet Mathematics*, 9(1):9–18, 1978. 4, 10

[SYN15] Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of neural nets through robust optimization. *arXiv preprint arXiv:1511.05432*, 2015. 2

[SZS+14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 1, 2, 4, 6

[TG16] Thomas Tanay and Lewis Griffin. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016. 2

[TTV16] Pedro Tabacof, Julia Tavares, and Eduardo Valle. Adversarial images for variational autoencoders. *arXiv preprint arXiv:1612.00155*, 2016. 2