How many dimensions are required to find an adversarial example?

Charles Godfrey,[1] Henry Kvinge,[1,2] Elise Bishoff,[1] Myles Mckay,[1,3]
Davis Brown,[1] Tim Doster,[1] and Eleanor Byler[1]
[1]Pacific Northwest National Laboratory, [2]Department of Mathematics, University of Washington,
[3]Department of Astronomy, University of Washington
{first}.{last}@pnnl.gov

Abstract

Past work exploring adversarial vulnerability have focused on situations where an adversary can perturb all dimensions of model input. On the other hand, a range of recent works consider the case where either (i) an adversary can perturb a limited number of input parameters or (ii) a subset of modalities in a multimodal problem. In both of these cases, adversarial examples are effectively constrained to a subspace V in the ambient input space $\mathcal{X}$. Motivated by this, in this work we investigate how adversarial vulnerability depends on $\dim(V)$. In particular, we show that the adversarial success of standard PGD attacks with ℓ^p norm constraints behaves like a monotonically increasing function of $\epsilon(\frac{\dim(V)}{\dim \mathcal{X}})^{\frac{1}{q}}$ where ϵ is the perturbation budget and $\frac{1}{p}+\frac{1}{q}=1$, provided $p>1$ (the case $p=1$ presents additional subtleties which we analyze in some detail). This functional form can be easily derived from a simple toy linear model, and as such our results land further credence to arguments that adversarial examples are endemic to locally linear models on high dimensional spaces.

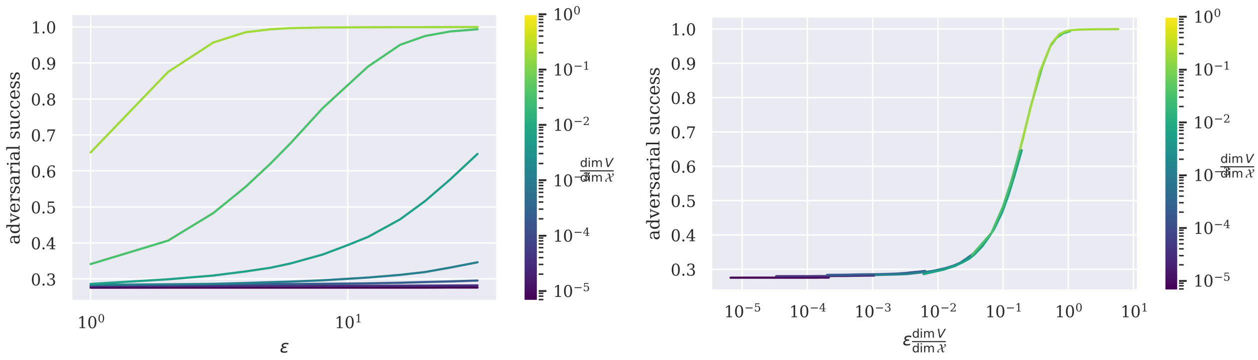
1. Introduction

Since they were first identified in [33], there has been strong sense that a model's vulnerability to adversarial examples is strongly connected to the dimension of its input space. This connection has been mined by a range of works which use it as a perspective with which to explain the prevalence of adversarial examples in certain model types (e.g., computer vision) – in Sec. 2 we provide a brief synopsis of this research. As deep learning models are applied to more and more safety critical applications, there is also an increasing practical relevance to understanding any general connections between adversarial vulnerability and the properties of a problem. In such settings, a simple statistic that can be easily computed (such as model input dimension) is useful for gauging the general adversarial risk for a proposed deep learning system.

This is especially true when the proposed system uses less familiar modalities/tasks to which one cannot easily refer to studies in the literature. For example, suppose one needs to evaluate the safety of applying deep learning to the output of a range of different sensors. Past work has considered the ambient dimension in which this data is collected. Should we worry less if a sensor captures a signal as a 50-dimensional vector rather than a $5,000$-dimensional vector? In this paper we take this line of reasoning a step further and ask how this situation changes when instead of changing the ambient dimension we change the dimension of the subspace in which one is constrained to perturb input. Such a thought experiment has practical relevance. Suppose that of the 500 input dimensions to our model, we believe that an adversary is only likely to get access to 50 dimensions (this may happen in multimodal settings where an adversary has much better access to a subset of the modalities). How should we compare this to a situation in which are only able to perturb a fixed 100-dimensional subspace of the input? How about a 5-dimensional subspace?

Motivated by this, in this work we revisit the connection between dimension and adversarial vulnerability. Unlike most other works in this space, which look at susceptibility to adversarial examples as a function of the number of input dimensions $\dim(\mathcal{X})$ alone, we explore model susceptibility to adversarial examples constrained to a subspace $V \subseteq \mathcal{X}$ as a function of $\dim(V)/\dim(\mathcal{X})$. We find that unsurprisingly, for fixed $\dim(\mathcal{X})$, as $\dim(V)$ decreases average adversarial success rate (ASR) also decreases, though ASR only drops significantly when the quotient $\dim(V)/\dim(\mathcal{X})$ drops below around 10% (see Figure 1). In other words, a model remains vulnerable when an adversary is only able to perturb a subset of input dimensions, but as this subset covers an ever smaller fraction of the available dimensions an adversary has to put increasing effort into finding adversarial examples.

We further study how the adversarial budget ϵ with respect to the ℓ^p-norm interacts with $\dim(V)$ and $\dim(\mathcal{X})$.



(a) (b)

Figure 1. **(a)** Success of PGD adversarial attacks on an ImageNet trained ResNet50, with ℓ^∞-norm constraints on perturbation budget, constrained to subspaces $V \subseteq \mathcal{X}$ spanned by $\dim V$ randomly selected standard basis vectors. Adversarial examples are computed for a random subsample of 10,000 datapoints from the ImageNet validations set. The x-axis is the ϵ-bound used during example generation and the different colored curves indicate the dimension $\dim V$ of the subspace to which the examples were constrained to, relative to the dimension $\dim \mathcal{X}$ $(= 3 \cdot 224^2)$ of the ambient space. When only a small number of dimensions can be perturbed, adversarial examples are challenging to generate even with large ϵ-bounds. **(b)** These curves become aligned when we reparameterize the x-axis by scaling by $\frac{\dim V}{\dim \mathcal{X}}$.

We find that the relationships of ϵ to ASR for different $\dim(V)$ are nearly identical up to scaling: more specifically, suppose that $C_{V_1} : \mathbb{R} \to [0, 1]$ and $C_{V_2} : \mathbb{R} \to [0, 1]$ map adversarial budget to ASR when adversarial examples are constrained to subspaces V_1 and V_2 respectively. We find that

$$C_1\left(\left(\frac{\dim(V_1)}{\dim(\mathcal{X})}\right)^{1/q} \epsilon\right) \approx C_1\left(\left(\frac{\dim(V_2)}{\dim(\mathcal{X})}\right)^{1/q} \epsilon\right) \qquad (1.1)$$

where q satisfies $1/p + 1/q = 1$. This points to a strong relationship between $\dim(V)$, p, and ϵ that to our knowledge is novel. It further tells us that risk from adversarial examples can be mitigated by either restricting the dimensions that data can be manipulated ($\dim V$) in or restricting the amount they can be manipulated before they are noticed (ϵ). This relationship is consistent across values of $\dim(V)/\dim(\mathcal{X})$: if one wanted to understand the risk of an adversary purturbing data in a 50-dimensional subspace of a 500-dimensional-input space, one could for example estimate the success rate of an adversary with access to the entire input space and extrapolate using Eq. (1.1). Finally, we provide a theoretical backing for our results as well as analyze their implications on common theories behind the prevalence of adversarial examples in Section 6.

In summary, our contributions in this paper include the following.

- We run a range of experiments restricting adversarial examples to a fixed subspace V of input space and explore how the dimension of V impacts adversarial success rate (ASR).
- Our results show that there are predictable trade-offs between ϵ and $\dim(V)$. That is, we can scale ϵ and $\dim(V)$ (dependent on the ℓ^p-norm used) so that ASR remains fixed.
- We provide a theoretical basis for our observations and analyze what this says about different theories explaining the existence of adversarial examples.

2. Related work

Why do adversarial examples exist?: There have been many proposed explanations for the phenomena of adversarial examples; we provide an incomplete but representative sample. A number of works such as [5, 29] present explanations in terms of dimensionality curses. In [14] it is argued that adversarial examples are a side-effect of locally linear behavior of deep learning models, an idea that is further investigated both theoretically and empirically in [6]. This theme also appears in [3, 4], which (among many other things) prove ReLU networks with multiple layers are linear on large regions of input space.

Adversarial examples and input dimension: A range of works have looked at the connection between the input dimension of data to a model and the prevalence of adversarial examples. Such works include [31], which simplifies the set-up by approximating neural networks with their gradients, hence reducing the problem to linear classifiers. They vary input dimension by up-sampling CIFAR10. [29] derives formulae relating adversarial vulnerability to model input dimension $\dim \mathcal{X}$, adversarial budget ϵ (in arbitrary ℓ^p norms, including $p = 0$) and notably properties of the data distribution, and carries out experiments varying input dimension by up-sampling MNIST. [5] includes theoretical results of a similar flavor, and also varies input dimension of image datasets by up-sampling as well as dimension-reducing preprocessing operations like the singular value decomposition. Unlike our work, none of these considered adversarial examples constrained to subspaces $V \subset \mathcal{X}$.[1]

Adversarial examples constrained to subspaces: There is a continually expanding body of work on adversarial perturbations constrained to submanifolds of the input space of a model. [15, 17, 28, 30, 38] all study the vulnerability of neural networks to perturbations constrained to a subspace corresponding to some Fourier frequency range (for instance, high, low or intermediate frequencies). [21, 39] study vulnerability to perturbations which modify color curves simultaneously at all locations of an image.

Among works most in line with the present one, [10] studies the minimal norm perturbation $\delta \in V \subseteq \mathcal{X}$ required to move an input $x \in \mathcal{X}$ across a decision boundary of f. They provided theoretical results for linear classifiers (and more general models in terms of curvature properties of decision boundaries) as well as empirical results for several image classifiers. The main theorems of [10] state that the norm of the minimal perturbation δ scales like $\sqrt{\frac{\dim \mathcal{X}}{\dim V}}$. However, they do not directly connect these findings with model error (a.k.a. adversarial success) and their analysis is limited to the ℓ^2 norm (hence their theorems do not contradict Fig. 1, which illustrates ℓ^∞ adversarial success). On the other hand, in this work we consider arbitrary ell_p-norm bounds and actually connect p to the rate of growth of adversarial success rate. The work in [10] is also intimately connected with the DeepFool attack [26],[2] as well as [9, 11]. In contrast, we mostly focus on PGD attacks due to their universality ([23]) and prevalence in the adversarial machine learning literature.

A number of works such as [18, 32, 35] ask the opposite of our question. Namely, what subspace $V \subset \mathcal{X}$ adversarial perturbations tend to lie in. A consistent finding of [18, 32] is that in situations where the data distribution lies on a manifold $M \subseteq \mathcal{X}$, adversarial examples for data points $x \in M$ tend to lie in the *normal space* NM_x, whose dimension is the codimension of M — [18] observes increasing vulnerability as this codimension increases.

Perhaps the work most similar to what we present here is [8], which investigates the phenomena of low-dimensional adversarial perturbations with theoretical results and empirical confirmations. Our findings are generally consistent with theirs, and we build on [8] with an extensive empirical analysis of *simultaneous* dependence of adversarial success on $\dim V$, ϵ and the metric under consideration (i.e. ℓ^2 or ℓ^∞). In addition, our experiments involve much larger models and datasets — we hope the description of our methods in Appendix B illustrates that this scaling-up is not trivial.

Finally we note that our results can be tied to a range of studies that look at statistics of adversarial examples with respect to different situations. For example, [6] studies scaling properties of adversarial success with respect to the perturbation constraint ϵ, obtaining results qualitatively similar to ours along that axis of variation.

3. Adversarial examples and subspaces

Let $f : \mathcal{X} \to \mathcal{Y}$ be a classification model, where $\mathcal{X}$ is the space of input data and $\mathcal{Y}$ the space of model outputs. We focus on image classifiers so that $\mathcal{X}$ is a space of pixel values (a hypercube of the form $[0, 1]^n$ for some n depending on image resolution) and $\mathcal{Y} = \{1, \dots, K\}$ with K the number of classes. The models f we consider are deep neural networks. By definition an **adversarial example** for a data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is a model input $x' = x + \delta \in \mathcal{X}$ such that

- $f(x) = y$ but $f(x') \neq y$ (x' is misclassified) and
- $d(x', x) < \epsilon$ where d is some chosen metric and $\epsilon > 0$ some chosen constraint (x' is close to x).

We will measure **adversarial success** as the probability that $f(x') \neq y$; this probability depends on the algorithm used to generate $x' = x + \delta \in \mathcal{X}$, and empirically can be estimated by generating perturbations for all x in a validation dataset and computing the error of f on the resulting "adversarial dataset."[3]

Standard methods of generating adversarial examples [14, 33] perturb model inputs by independently modifying all pixel values, however as early as [14] it was observed that sparse perturbations modifying only a subset of pixel values were also effective. By now there exist a plethora of adversarial example generation techniques which optimize for perturbations δ constrained to a subspace[4] $V \subseteq \mathcal{X}$, in many cases with $\dim V$ a small fraction of $\dim \mathcal{X}$ — a common aim of these methods is to modify x in a way that is perceptually natural (so that x' will appear innocuous even to a human-in-the-loop) while using relatively few parameters. We discuss a representative sample of such techniques in Sec. 2. Such widespread interest in constrained adversarial perturbations $\delta \in V \subset \mathcal{X}$ raises a foundational question:

$$\textit{how does adversarial success depend on } \frac{\dim V}{\dim \mathcal{X}}\textit{?} \quad (3.1)$$

In Sec. 4 we design experiments to measure this dependency

[1] At first glance it might seem the SVD preprocessing lands in a proper subspace $V \subset \mathcal{X}$, but it is more accurate to say it decreases the ambient dimension $\dim \mathcal{X}$.

[2] Which does implicitly indicate the correct scaling for more general ℓ^p-metrics.

[3] In particular, we do not require the model to correctly classify the unperturbed input, i.e. we do not restrict attention to data points where $f(x) = y$.

[4] Sometimes, but not always, an affine linear subspace.

for a variety of families of subspaces $V \subset \mathcal{X}$ (including those spanned by random subsets of pixels or random sets of orthogonal vectors) and metrics (including ℓ^2 and ℓ^∞).[5] We repeatedly find that success of adversarial attacks constrained in the ℓ^p metric is a function of $\epsilon \cdot (\frac{\dim V}{\dim \mathcal{X}})^{\frac{1}{q}}$, where $\frac{1}{p} + \frac{1}{q} = 1$ — this is illustrated in Fig. 1 and described further in Sec. 5.2.

This experiment serves as a lens through which to investigate two common, not-necessarily-mutually-exclusive explanations for the prevalence of adversarial examples in deep learning — these are:

(i) **Adversarial examples are a result of the curse of dimensionality:** a deep learning model $f : \mathcal{X} \to \mathcal{Y}$ subdivides the high dimensional input space $\mathcal{X}$ into decision regions $f^{-1}(y) \subseteq \mathcal{X}$. A variety of well-studied toy models, such as binary linear classification of points on a sphere $S^{n-1} \subset \mathbb{R}^n$, have the property that in high dimensions every $x \in f^{-1}(y)$ lies very close to the boundary of $f^{-1}(y)$ (for linear classification of points on S^{n-1} this is the statement that "as $n \to \infty$, all the volume lies near the equator").

(ii) **Adversarial examples are a consequence of (locally) linear behavior:** At least locally, f is well approximated by an affine linear function $Wx + b$, and for appropriately chosen δ we can make $|W\delta|$ large enough to ensure $f(x + \delta) \neq f(x)$.

These two explanations are more closely related than they might initially appear. For example, in Item (ii) the number of coefficients of w equals $\dim \mathcal{X}$, and as pointed out in [14] the fast gradient sign method exploits the fact that when $\delta = -\operatorname{sign}(w)$, $w^T\delta = -\sum_i |w_i| = -|w|_1$ which scales with $\dim \mathcal{X}$ (provided the scale of the coefficients $w_1, \ldots, w_{\dim \mathcal{X}}$ is fixed). Here the idea is that the number of terms in the sum $\sum_i |w_i|$ is $\dim \mathcal{X}$, so if the coefficients w_i are IID $E[\sum_i |w_i|] = \dim \mathcal{X} \cdot E[|w_1|]$ — in some sense, this is also a curse of dimensionality. In Sec. 6 we compare the various *theoretically predicted* scaling properties of adversarial success with respect to ϵ and $\dim V$, lifted from papers arguing for Item (i) or Item (ii).

4. Perturbations in random subspaces

Designing an experiment to measure adversarial success with varying $\dim V$ and ϵ requires making a number of choices:

(i) a distribution of subspaces $V \subseteq \mathcal{X}$ to sample from, or more technically speaking a probability distribution on a Grassmannian $\mathrm{Gr}(\dim V, \dim \mathcal{X})$,

(ii) a metric d used to define constraints on perturbations, and

(iii) an adversarial example generation algorithm $\mathcal{A}$.

[5] By definition for any $p \geq 1$ the ℓ^p-distance between points $x, y \in \mathbb{R}^n$ is $(\sum_i |x_i - y_i|^p)^{\frac{1}{p}}$.

To establish a baseline, we consider the case where the distribution of subspaces is either *uniform* or the distribution obtained by taking V to be the span of $\dim V$ standard basis vectors $e_i \in \mathcal{X}$ sampled uniformly. We restrict attention to the ℓ^p metrics for $p \in \{2, \infty\}$, and look at adversarial examples generated by projected gradient descent as in [23]. We also must specify (i) a dataset $\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}$ of images, and (ii) an image classifer $f : \mathcal{X} \to \mathcal{Y}$.

Having made these decisions, for a fixed dimension d and constraint ϵ and for each data point $(x, y) \in \mathcal{D}$, we sample a d-dimensional subspace $V \subseteq \mathcal{X}$ according to the specified distribution on $\mathrm{Gr}(d, \dim \mathcal{X})$, generate an adversarial perturbation $x' \in \mathcal{X}$ constrained to V and with $d(x', x) \leq \epsilon$ using the algorithm $\mathcal{A}$, and record whether the attack was successful, that is: $\mathbf{1}(f(x') \neq y)$. To obtain a low-variance estimate of adversarial success we average $\mathbf{1}(f(x') \neq y)$ over the dataset $\mathcal{D}$ (or a reasonably large subsample thereof) and sample a different subspace $V \subseteq \mathcal{X}$ for each datapoint (x, y) to approximate

$$\mathrm{success}(d, \epsilon) = P(f(x') \neq y). \tag{4.1}$$

It should be emphasized that we are computing statistics for random subspaces; as other works discussed in Sec. 2 have shown, there are specific subspaces in which a higher adversarial success rate can likely be achieved.

5. Experiments

5.1. Datasets and models

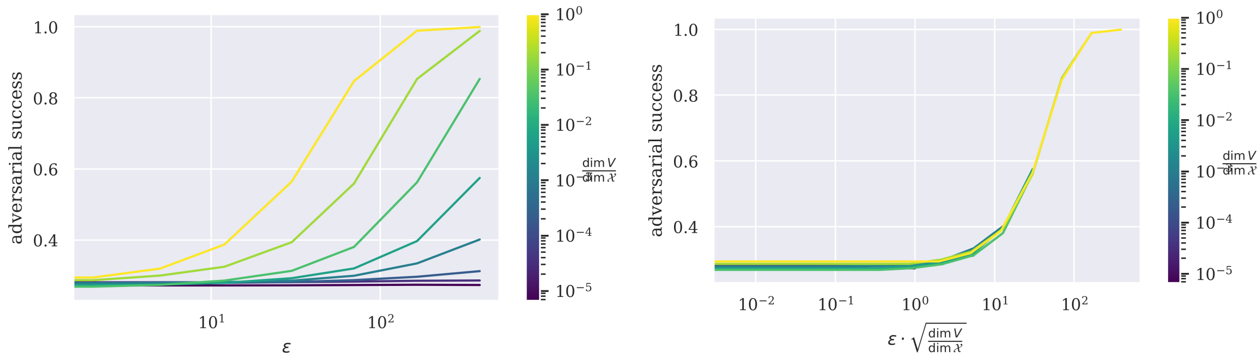
We experiment with several image classification data sets and model architectures, of increasing image resolution and network capacity:

- The small convolutional network used in [23] trained on the MNIST dataset [22].
- A ResNet9 [16] trained on the CIFAR10 dataset [19].
- A ResNet50 [16] trained on the ImageNet dataset [7].

For further details on model architectures and training, we refer to Appendix B.1.

5.2. Functional form of adversarial success

We may view the adversarial successes $\mathrm{success}(\dim V, \epsilon)$ as a sequence of functions of ϵ, one for each $\dim V \in \{1, \ldots, \dim \mathcal{X}\}$ as shown in the top plot of Fig. 1, which displays results of ℓ^∞ PGD adversarial attacks constrained to spans of subsets of standard basis vectors for a ResNet50 trained on ImageNet. It appears that for varying $\dim V$, the curves $\mathrm{success}(\dim V, \epsilon)$ differ by x-axis scalings, that is, transformations of the form $\mathrm{success}(\dim V, \epsilon) \leftarrow \mathrm{success}(\dim V, \lambda\epsilon)$ for some $\lambda > 0$. This is indeed the case: the bottom plot in Fig. 1 shows that the curves $\mathrm{success}(\dim V, \epsilon \cdot \frac{\dim V}{\dim \mathcal{X}})$ are almost identical. Figures 5 and 6 show the same phenomena for a 2-layer CNN on MNIST and a ResNet9 on CIFAR10.



(a) (b)

Figure 2. **(a)** Success of PGD adversarial attacks on an ImageNet trained ResNet50, with ℓ^2-norm constraints on perturbation budget, constrained to subspaces $V \subseteq \mathcal{X}$ spanned by $\dim V$ randomly selected standard basis vectors. Adversarial examples are computed for a random subsample of 10,000 datapoints from the ImageNet validations set. The x-axis is the ϵ-bound used during example generation and the different colored curves indicate the dimension $\dim V$ of the subspace to which the examples were constrained to, relative to the dimension $\dim \mathcal{X}$ $(= 3 \cdot 224^2)$ of the ambient space. When only a small number of dimensions can be perturbed, adversarial examples are challenging to generate even with large ϵ-bounds. **(b)** These curves become aligned when we reparameterize the x-axis by scaling by $\sqrt{\frac{\dim V}{\dim \mathcal{X}}}$.

We can think of this analysis as expressing a *decomposition* $\text{success}(\dim V, \epsilon) = g(\epsilon \cdot \frac{\dim V}{\dim \mathcal{X}})$ into a composition of two functions, the first being the map $(\dim V, \epsilon) \mapsto \epsilon \cdot \frac{\dim V}{\dim \mathcal{X}}$, the second map g being some single-variable function applied to $\epsilon \cdot \frac{\dim V}{\dim \mathcal{X}}$. We do not attempt to identify this g.

Figure 2 shows analogous results for PGD adversarial examples constrained in the ℓ^2-norm (with plots for other architecture types and models in Figure 7 and Figure 8). In this case, the reparametrization of the x-axis that results in almost identical curves is obtained by replacing $\text{success}(\dim V, \epsilon)$ with $\text{success}(\dim V, \epsilon \cdot \sqrt{\frac{\dim V}{\dim \mathcal{X}}})$. This shows that the functional form of adversarial success in terms of ϵ and $\dim V$ depends on the norm constraining adversarial perturbations. In the following section, we argue that as long as $p > 1$ adversarial success with ℓ^p-constraints depends on $\epsilon d^{\frac{1}{q}}$ where $\frac{1}{p} + \frac{1}{q} = 1$, a hypothesis consistent with experimental results in Fig. 2 and Fig. 1.

The case $p = 1$ is more complicated, and we defer its analysis to Appendix B.5.

6. Comparison with existing theoretical predictions

There are many existing works investigating the mathematical source of adversarial examples for deep learning models. Several of these include (either as a main result or a byproduct of calculations) predictions for the functional form of adversarial success in terms of perturbation budget ϵ and the dimension $\dim V$ to which perturbations are constrained. We reviewed a subsample of such papers. Note that most of these papers (the exceptions being [8, 10]) focus on the dimension of the input space of a model alone and do not consider the additional constraint that adversarial examples are confined to a subspace.

(i) From the analysis in [14] one can predict that adversarial success is a function of $\epsilon \frac{\dim V}{\dim X}$ (for $p = \infty$).

(ii) [6, 8, 10] predict that adversarial success is a function of $\epsilon (\frac{\dim V}{\dim \mathcal{X}})^{\frac{1}{2}}$ (for $p = 2$).

(iii) From the analysis in [26,31] one can predict that adversarial success is a function of $\epsilon (\frac{\dim V}{\dim \mathcal{X}})^{\frac{1}{q}}$ (for $\frac{1}{p} + \frac{1}{q} = 1$).

(iv) [29] predicts that adversarial success is a function of $\epsilon \dim V^{\frac{1}{2} - \frac{1}{\min(2,p)}}$.

The predictions of Items (i) to (iii) are all consistent with our experimental results, suggesting that the situation where adversarial examples are constrained to a subspace of dimension d is effectively equivalent to the unconstrained situation where data is found in an ambient space of dimension d. Those of Item (iv) are not obviously consistent with our experimental data, although we refrain from saying that they are inconsistent since the analysis of [29] involves a series of inequalities, and it is unclear how the predictions would change if one used slightly different approximations.[6]

The dependence of adversarial success on $\epsilon (\frac{\dim V}{\dim \mathcal{X}})^{\frac{1}{q}}$ can be derived from the simplest possible toy model, namely binary linear classification. Let $\mathcal{X} = \mathbb{R}^n$ and suppose

$$f(x) = w^T x + b, \text{ for some } w \in \mathbb{R}^n, b \in \mathbb{R}. \quad (6.1)$$

[6] Moreover, the aim of [29] was to demonstrate prevalence of adversarial examples, not to estimate functional forms for adversarial success.

Let $V \subseteq \mathbb{R}^n$ be a subspace with $\dim V = d$. We will assume there exists an isometry $U : \mathbb{R}^d \xrightarrow{\simeq} V$ with respect to the ℓ^p metric — note that this puts strong restrictions on U and V when $p \neq 2$ (in that case, V must be spanned by standard basis vectors and U must be an orthogonal matrix with entries in $\{-1, 0, 1\}$). The point x admits an ℓ^p adversarial example in the subspace V with budget ϵ, i.e. there is a $\delta \in V$ such that $\operatorname{sign} f(x + \delta) \neq \operatorname{sign} f(x)$ and $|\delta|_p \leq \epsilon$, if and only if the ℓ^p *margin of* x

$$\min_{\delta \in V} \{|\delta|_p \mid f(x+\delta) = 0\} \tag{6.2}$$

is at most ϵ.

Lemma 6.3. *With the above definitions and notation and with* $\frac{1}{p} + \frac{1}{q} = 1$,

$$\min_{\delta \in V} \{|\delta|_p \mid f(x+\delta) = 0\} = \frac{|w^T x + b|}{|w^T U|_q}. \tag{6.4}$$

Appendix A contains a proof. Our experimental results only measure the *probability* that x admits an ℓ^p-adversarial example in the subspace V with budget ϵ. By the above lemma, this probability is $P(\frac{|w^T x+b|}{|w^T U|_q} \leq \epsilon)$, which can be rewritten as $P(|w^T x + b| \leq \epsilon |w^T U|_q)$. We claim that when $p > 1$ and V (equivalently U) is sampled with sufficient randomness

$$\mathbb{E}\big[|w^T U|_q\big] = \Big(\frac{d}{n}\Big)^{\frac{1}{q}} |w|_q. \tag{6.5}$$

In the case where V is generated by a subset, say $\{e_{i_1}, \ldots, e_{i_d}\}$ of basis vectors, this can be argued as follows:

$$\begin{aligned} \frac{|w^T U|_q^q}{|w|_q^q} &= \frac{\sum_{j=1}^d |w_{i_j}|^q}{\sum_{i=1}^n |w_i|^q} \\ &= \frac{d}{n} \frac{\frac{1}{d}\sum_{j=1}^d |w_{i_j}|^q}{\frac{1}{n}\sum_{i=1}^n |w_i|^q} \end{aligned} \tag{6.6}$$

When the basis subset $\{e_{i_1}, \ldots, e_{i_d}\}$ is sampled uniformly[7] we claim that that the expectation of the term $\frac{1}{d}\sum_{j=1}^d |w_{i_j}|^q$ is exactly $\frac{1}{n}\sum_{i=1}^n |w_i|^q$ (at least when $d = 1$ this is immediate). Thus after averaging over many random subspaces V,

$$\mathbb{E}\Big[\frac{\frac{1}{d}\sum_{j=1}^d |w_{i_j}|^q}{\frac{1}{n}\sum_{i=1}^n |w_i|^q}\Big] = 1, \text{ hence } \mathbb{E}\Big[\frac{|w^T U|_q^q}{|w|_q^q}\Big] = \frac{d}{n}. \tag{6.7}$$

Taking q-th roots and rearranging gives Eq. (6.5).

Note that in our experiments we compute something analogous to $P(|w^T x + b| \leq \epsilon |w^T U|_q)$ where probability is with respect to the underlying distribution of x and choice of U. Using a "point estimate" and replacing $|w^T U|_q$ with its mean $(\frac{d}{n})^{\frac{1}{q}} |w|_q$, one would simplify to $P(|w^T x + b| \leq \epsilon (\frac{d}{n})^{\frac{1}{q}} |w|_q)$, which since we treat w, b and the distribution of x as given is a function of $\epsilon (\frac{d}{n})^{\frac{1}{q}}$.

When $p = 1$, so $q = \infty$, Lemma 6.3 remains valid but the tricks applied in Eqs. (6.6) and (6.7) do not make sense, and indeed our experimental results in Appendix B.5 suggest dependence of adversarial success on $\epsilon (\frac{\dim V}{\dim \mathcal{X}})^{\frac{1}{q}}$ alone breaks down somewhat in this case. For further analysis of this case, we refer to Appendix B.5.

[7]For example, by taking $i_1, \ldots, i_d = \sigma(1), \ldots, \sigma(d)$ where σ is a uniformly random permutation of $\{1, \ldots, n\}$.

7. Limitations

Adversarial examples given by gradient-based perturbations with ℓ^p constraints make up only one (and arguably, a narrow) type of distribution-shifted test data causing machine learning model failure. For further discussion of this point see [12, 13]. While we take inspiration from adversarial example generators constraining perturbations to subspaces (surveyed in Sec. 2), our experiments are limited to the baseline of random subspace selection (whereas most subspace-constrained adversarial example generators choose their subspace more carefully). We also only experiment with image classifiers, though adversarial examples have been found to exist for essentially all deep learning systems [20, 25, 37].

8. Conclusion and open questions

We demonstrate that the adversarial success $\operatorname{success}(\dim V, \epsilon)$ of PGD attacks constrained to a (random) $\dim V$-dimensional subspace V of the model input space $\mathcal{X}$ with ℓ^p budget ϵ (and $p > 1$) is essentially a function of the single variable $\epsilon (\frac{\dim V}{\dim \mathcal{X}})^{\frac{1}{q}}$ where $\frac{1}{p} + \frac{1}{q} = 1$ (rather than a function of two variables as considered in prior work). The fact that this relationship can be derived in the toy example of a linear binary classifier, and holds quite sharply in all our experiments, seems to lend further credence to the theory that adversarial examples are a byproduct of the locally linear behavior of neural networks with high dimensional input spaces.

Acknowledgments

The research described in this paper was conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy.

References

[1] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with Sparsity-Inducing Penalties, Nov. 2011. 9

[2] R. R. Bahadur. A Note on Quantiles in Large Samples. *The Annals of Mathematical Statistics*, 37(3):577–580, June 1966. 10, 11

[3] Peter Bartlett, Sebastien Bubeck, and Yeshwanth Cherapanamjeri. Adversarial Examples in Multi-Layer Random ReLU Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9241–9252. Curran Associates, Inc., 2021. 2

[4] Sebastien Bubeck, Yeshwanth Cherapanamjeri, Gauthier Gidel, and Remi Tachet des Combes. A single gradient step finds adversarial examples on random two-layers neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 10081–10091. Curran Associates, Inc., 2021. 2

[5] Nandish Chattopadhyay, Anupam Chattopadhyay, Sourav Sen Gupta, and Michael Kasper. Curse of dimensionality in adversarial examples. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. 2

[6] Ekin D. Cubuk, Barret Zoph, Samuel S. Schoenholz, and Quoc V. Le. Intriguing Properties of Adversarial Examples, Nov. 2017. 2, 3, 5

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR09*, 2009. 4

[8] Elvis Dohmatob, Chuan Guo, and Morgane Goibert. Origins of low-dimensional adversarial perturbations, 2022. 3, 5

[9] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Mach. Learn.*, 107(3):481–508, 2018. 3

[10] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 3, 5

[11] Jean-Yves Franceschi, Alhussein Fawzi, and Omar Fawzi. Robustness of classifiers to uniform ℓ_p and gaussian noise. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1280–1288. PMLR, 09–11 Apr 2018. 3

[12] Justin Gilmer, Ryan P. Adams, Ian Goodfellow, David Andersen, and George E. Dahl. Motivating the Rules of the Game for Adversarial Example Research, July 2018. 6

[13] Justin Gilmer and Dan Hendrycks. A Discussion of 'Adversarial Examples Are Not Bugs, They Are Features': Adversarial Example Researchers Need to Expand What is Meant by 'Robustness'. *Distill*, 4(8):e00019.1, Aug. 2019. 6

[14] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 2, 3, 4, 5

[15] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low Frequency Adversarial Perturbation. In *UAI*. 3

[16] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[17] Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *ArXiv*, abs/1711.11561, 2017. 3

[18] Marc Khoury and Dylan Hadfield-Menell. On the geometry of adversarial examples. 2018. 3

[19] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 4

[20] Aditya Kuppa, Slawomir Grzonkowski, Muhammad Rizwan Asghar, and Nhien-An Le-Khac. Black box attacks on deep anomaly detectors. In *Proceedings of the 14th international conference on availability, reliability and security*, pages 1–10, 2019. 6

[21] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10408–10418, 2019. 3

[22] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 1998. 4

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 3, 4, 9, 10, 11

[24] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery. 9

[25] Natalie Maus, Patrick Chao, Eric Wong, and Jacob Gardner. Adversarial prompting for black box foundation models. *arXiv preprint arXiv:2302.04237*, 2023. 6

[26] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016. 3, 5

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Informa-*

tion Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019. 9

[28] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 09–15 Jun 2019. 3

[29] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2019. 2, 5

[30] Yash Sharma, G. Ding, and Marcus A. Brubaker. On the Effectiveness of Low Frequency Perturbations. 3

[31] Carl-Johann Simon-Gabriel, Yann Ollivier, Leon Bottou, Bernhard Schölkopf, and David Lopez-Paz. First-order adversarial vulnerability of neural networks and input dimension. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5809–5817. PMLR, 09–15 Jun 2019. 2, 5

[32] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. 2019. 3

[33] 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. 1, 3

[34] The Mosaic ML Team. composer. `https://github.com/mosaicml/composer/`, 2021. 9

[35] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017. 3

[36] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. 10

[37] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017. 6

[38] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *NeurIPS*, pages 13255–13265, 2019. 3

[39] Zhengyu Zhao, Zhuoran Liu, and Martha A. Larson. Adversarial color enhancement: Generating unrestricted adversarial images by optimizing a color filter. In *BMVC*, 2020. 3