ISSN 1346-5597

NII National Institute of Informatics

NII Technical Report

The Vulnerability of Learning to Adversarial Perturbation Increases with Intrinsic Dimensionality

Laurent Amsaleg, James Bailey, Sarah Erfani, Teddy Furon, Michael E. Houle, Miloš Radovanović, Nguyen Xuan Vinh

NII-2016-005E
June 2016

URL: http://www.nii.ac.jp/ Tokyo 101-8430, JAPAN +81-3-4212-2000 FAX: +81-3-3556-1916

The Vulnerability of Learning to Adversarial Perturbation Increases with Intrinsic Dimensionality

Laurent Amsaleg[1] James Bailey[2]
Sarah Erfani[2] Teddy Furon[1]
Michael E. Houle[3] Miloš Radovanović[4]
Nguyen Xuan Vinh[2]

[1]Equipe LINKMEDIA, INRIA/IRISA Rennes, France
[2]Dept. of Computing and Information Systems,
The University of Melbourne, Australia
[3]National Institute of Informatics, Tokyo, Japan
[4]Dept. of Mathematics and Informatics,
University of Novi Sad, Serbia
Correspondence email: meh@nii.ac.jp

Abstract

Recent research has shown that machine learning systems, including state-of-the-art deep neural networks, are vulnerable to adversarial attacks. By adding to the input object an imperceptible amount of adversarial noise, it is highly likely that the classifier can be tricked into assigning the modified object to any desired class. Furthermore, these adversarial samples generalize well across models: samples generated using one network can often succeed in fooling other networks or machine learning models. These alarming properties of adversarial samples have drawn increasing interest recently, with several researchers having attributed the adversarial effect to different factors, such as the high dimensionality of the data or the overly-linear nature of modern neural networks. Nevertheless, a complete picture of the cause of adversarial samples has not yet emerged. Towards this goal, we present a novel theoretical result that formally links the adversarial vulnerability of learning to the intrinsic dimensionality of the data. In particular, our investigation formally establishes that as the local intrinsic dimensionality (LID) increases, 1-NN classifiers become increasingly prone to being subverted. We show that in expectation, a k-nearest neighbor of a test point can be transformed into its 1-nearest neighbor by adding an amount of noise that diminishes as the LID increases. We also provide an experimental validation of the impact of LID on adversarial perturbation for both synthetic and real data, and discuss the implications of our result for general classifiers.

1 Introduction

Recent research has shown that the performance of machine learning systems, including state-of-the-art deep neural networks, can be subverted by a form of adversarial attack, in which a small amount of carefully-designed, imperceptible adversarial noise is added to an input object so as to influence a classification result [19, 26]. Moreover, it is often possible to cause a well-performing image classifier to misclassify a visually-identical test image to any other desired class, by engineering a suitable perturbation (see Figure 1). Adversarial perturbation generalizes surprisingly well across different models: an adversarial sample designed using one model can often succeed in subverting other models [26]. These alarming properties of adversarial perturbation carry many practical implications in an era where machine learning technologies are ubiquitous.

Adversarial attacks on learning systems can potentially cause tremendous damage and disruption. For example, an adversary could conceivably create a false passport or other identification in which the image appears to match his or her own face, but is recognized by the system as belonging to another individual (presumably one whose identity has been stolen). Another example is that of autonomous systems such as those found in surveillance systems or self-driving vehicles, which presumably rely heavily on machine learning technology for scene recognition. An adversarial attack could result in a disastrously erroneous decision by the system, even under the close scrutiny of a human overseer (such as a security official or automobile passenger). Such situations can seriously undermine the safety of autonomous systems, as well as public confidence in them. Thus, the increasing social reliance on intelligent autonomous systems has recently sparked significant research effort into the understanding and prevention of adversarial attacks on classification.

Recent research has attempted to explain adversarial perturbation from different perspectives. At first glance, it is tempting to hypothesize that vulnerability to adversarial perturbation is a peculiarity of individual learning systems such as deep neural networks, with the effect being the result of a complex interplay between the model and the data. It was also initially thought that the effect is a consequence of overfitting; however, as adversarial samples tend to generalize well even across models of different types [26], this is not a likely explanation. Moreover, it has recently been shown that even models with parameters picked at random are unstable with respect to adversarial perturbation [23]. In [9], Goodfellow et al. conjectured that modern deep neural networks, particularly those built with rectified linear units, are vulnerable to adversarial perturbation due to their highly linear nature. Their vulnerability has also been attributed to the high dimensionality of the input space: when accumulated over many dimensions, minor changes can 'snowball' into large changes in the transfer function [9]. Despite the many hypotheses that have been posed in the literature, a complete picture on the causes of the adversarial perturbation effect is yet to emerge.

Towards this goal, in this paper we present, to the best of our knowledge, the first theoretical explanation of the adversarial effect for classification of objects, in terms of the LID model of local intrinsic dimensionity (ID) [1, 12]. In the context of Euclidean spaces, a constructive proof is provided within which any reference point can be perturbed so as to change a targeted k-nearest neighbor (k-NN) into a 1-nearest neighbor (1-NN). Since the argument works with distributions of points and not fixed point sets per se, the notion of neighbor is stated in terms of mathematical expectation:

with respect to a sample size n, a target location $\mathbf{z}$ is a k-NN of a reference point $\mathbf{x}$ *by expectation* if k out of the n sample points would be expected to lie within distance $d(\mathbf{x}, \mathbf{z})$ of $\mathbf{x}$. The result gives a method of construction of a perturbed point $\mathbf{y}$ for which $\mathbf{z}$ becomes a 1-NN of $\mathbf{y}$ by expectation, as n tends to infinity. Conditions on $\mathbf{y}$ are provided for a relationship to hold between the amount of perturbation required on the one hand, to the intrinsic dimensionality of the distance distribution from $\mathbf{x}$ on the other. The effect is such that as the intrinsic dimensionality at $\mathbf{x}$ rises, the amount of perturbation required tends to zero.

The remainder of the paper is organized as follows. In Section 2, we give a brief overview of adversarial perturbation and the concept of intrinsic dimensionality, together with a brief review of some of the useful properties of the LID model. In Section 3, we give a proof of our main theoretical result, followed in Section 4 by an experimental validation of the impact of intrinsic dimensionality on the adversarial perturbation effect. Section 5 concludes the paper with a discussion of some of the possible implications of our result for deep neural networks and other state-of-the-art learning systems.

2 Background

2.1 Adversarial perturbation

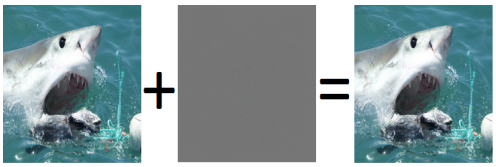
For a general machine learning model, adversarial perturbation can be designed as follows. Following the notation in [27], let $\mathbf{p} = f(\mathbf{x})$ be a classifier that, for each input object $\mathbf{x} \in \mathbb{R}^d$, outputs a vector of probabilities $\mathbf{p} = [p_1, \ldots, p_C]$ of the object belonging to each of the C predefined classes. We wish to add a small distortion $\mathbf{d} \in \mathbb{R}^d$ to $\mathbf{x}$, such that $f(\mathbf{x} + \mathbf{d})$ is close to a target adversarial probability $\mathbf{p}^A = [\mathbb{1}_{i=a}]$, which assigns zero probability to all but a chosen adversarial label a. One way to craft the adversarial noise $\mathbf{d}$ is by solving the following optimization problem:

$$\min_{\mathbf{d}} \ \|\mathbf{d}\| + \alpha D_{KL}(\mathbf{p}^A \| f(\mathbf{x} + \mathbf{d})), \ \text{subject to: } \mathbf{l} \leq \mathbf{x} + \mathbf{d} \leq \mathbf{u} \tag{1}$$

Here, $D_{KL}(\cdot)$ is the Kullback-Leibler divergence, $\mathbf{l}$ and $\mathbf{u}$ define the lower and upper bounds of the input domain respectively, and α is a balancing factor that determines the tradeoff between the level of distortion and the closeness to the target adversarial class label. With classifiers trained using gradient descent, the above optimization problem can be solved straightforwardly, using either gradient descent or box-constrained L-BFGS [26]. An example of an adversarial attack using this strategy is shown in Figure 1(a).

In this paper, we prove strong theoretical statements concerning the effect of perturbation on 1-NN classifiers. 1-NN classification has long been known to be 'asymptotically optimal', in that the probability of error is bounded above by twice the Bayes minimum probability of error, as the training set size tends to infinity [4, 25]. In this sense, an infinite sample set can be regarded as containing half the classification information in the nearest neighbor.

Within a Euclidean space or other vector space, 1-NN classification admits a relatively straightforward perturbation strategy that is particularly amenable to theoretical analysis. In order to transform a test point so that it is misclassified as a given target class, it is sufficient to select a point from the target class (presumably but not



(a) Using the optimization formulation in (1), adding adversarial noise (center) to an example of class 'Shark' (left) induces a deep neural network to assign the perturbed result (right) to the class 'Binocular'.

(b) Perturbation directly towards a target image. Although it is almost indistinguishable from the original image (left), the perturbed image (right) risks being assigned to the same class as the target image (center) by 1-NN classifiers.

Figure 1: Examples of adversarial perturbation.

necessarily the candidate closest to the test point), and perturb the test point toward the target point along the straight line joining them (for an example, see Figure 1(b)). Assuming that all data points are distinct, as the amount of perturbation increases, the perturbed point would eventually find itself with the target point as its 1-NN. Even for deep neural networks and other state-of-the-art classifiers of continuously-distributed data, it should be noted that a sufficiently-large perturbation directly towards a target point must eventually result in the test point entering a region associated with the class to which the target belongs.

2.2 Intrinsic dimensionality

Over the past decades, many characterizations of the ID of sets have been proposed: classical measures (mainly of theoretical interest), including the Hausdorff dimension, Minkowski-Bouligand or 'box counting' dimension, and packing dimension (for a general reference, see [5]); the correlation dimension [10]; 'fractal' measures of the space-filling capacity or self-similarity of the data [3, 6, 11]; topological estimation of the basis dimension of the tangent space of a data manifold from local samples [2, 7, 20, 21, 28]. Projection-based learning methods such as PCA [7] can produce as a byproduct an estimate of ID. Expansion-based models include the expansion dimension (ED) [17], the generalized expansion dimension (GED) [14], and the minimum neighbor distance (MiND) [21].

As a motivating example from m-dimensional Euclidean space, consider the situation which the volumes V_1 and V_2 are known for two balls of differing radii r_1 and r_2,

respectively, centered at a common reference point. The dimension m can be deduced from the ratios of the volumes and the distances to the reference point, as follows:

$$\frac{V_2}{V_1} = \left(\frac{r_2}{r_1}\right)^m \implies m = \frac{\ln V_2 - \ln V_1}{\ln r_2 - \ln r_1} .$$

For finite data sets, GED formulations are obtained by estimating the volume of balls by the numbers of points they enclose [14, 17].

Instead of regarding intrinsic dimensionality as a characteristic of a collection of data points (as evidenced by their distances from a supplied reference location), the GED was recently extended to a statistical setting, in which the distribution of distances to a query point is modeled as a continuous random variable $\mathbf{X}$ [1, 12]. The notion of volume is naturally analogous to that of probability measure. ID can then be modeled as a function of distance $\mathbf{X} = r$, by letting the radii of the two balls be $r_1 = r$ and $r_2 = (1+\epsilon)r$, and letting $\epsilon \to 0^+$.

Definition 1 ([12]) *Let $\mathbf{X}$ be a random distance variable. For any r such that $F_{\mathbf{X}}(r) > 0$, the* local intrinsic dimensionality *of $\mathbf{X}$ at r is given by*

$$\mathrm{ID}_{F_{\mathbf{X}}}(r) \triangleq \lim_{\epsilon \to 0^+} \frac{\ln F_{\mathbf{X}}((1+\epsilon)r) - \ln F_{\mathbf{X}}(r)}{\ln((1+\epsilon)r) - \ln r} = \frac{r \cdot F'_{\mathbf{X}}(r)}{F_{\mathbf{X}}(r)},$$

wherever the limit exists. The second equality follows by applying l'Hôpital's rule to the limits provided that $F_{\mathbf{X}}$ is positive and differentiable over an open interval containing r.

Under this distributional interpretation, the original data set determines a sample of distances from a given point. The intrinsic dimensionality (here referred to simply as 'local ID', or 'LID') of this distance distribution $F_{\mathbf{X}}$ is estimated. The definition of $\mathrm{ID}_{F_{\mathbf{X}}}$ can be extended to the case where $r = 0$, by taking the limit of $\mathrm{ID}_{F_{\mathbf{X}}}(r)$ as $r \to 0^+$, whenever this limit exists:

$$\mathrm{ID}_{F_{\mathbf{X}}}(0) \triangleq \lim_{r \to 0^+} \mathrm{ID}_{F_{\mathbf{X}}}(r) .$$

For an illustration of the intrinsic dimensionality of distance distributions, see Figure 2.

The smallest distances from a given point can be regarded as 'extreme events' associated with the lower tail of the underlying distribution. The modeling of neighborhood distance values can thus be investigated from the viewpoint of extreme value theory (EVT). In [13], it is shown that the EVT representation of the distance distribution $F_{\mathbf{X}}$ completely determines function $\mathrm{ID}_{F_{\mathbf{X}}}$, and that the EVT index is in fact identical to $\mathrm{ID}_{F_{\mathbf{X}}}(0)$.

Theorem 1 ([13]) *Let $F : (0, z) \to \mathbb{R}$ be a function over the range $(0, z)$, for some choice of $z > 0$ (possibly infinite). Let $v \in [0, z)$ be a value for which $\mathrm{ID}_F(v)$ exists. Then for any $r, w \in (0, z)$ such that F is positive and differentiable everywhere over an open interval containing $[\min\{r, w\}, \max\{r, w\}]$,*

$$\frac{F(r)}{F(w)} = \left(\frac{r}{w}\right)^{\mathrm{ID}_F(v)} \cdot G_{F,v,w}(r), \text{ where}$$

$$G_{F,v,w}(r) \triangleq \exp\left(\int_r^w \frac{\mathrm{ID}_F(v) - \mathrm{ID}_F(t)}{t}\, dt\right) .$$

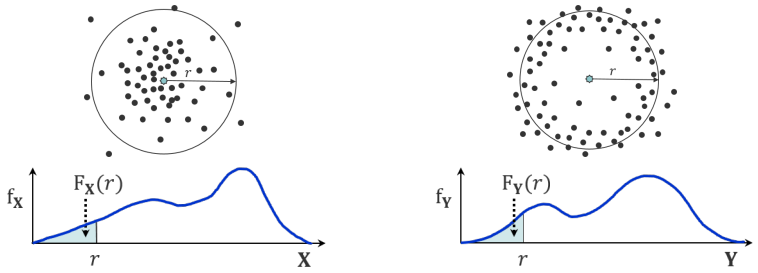


Figure 2: The random distance variables $\mathbf{X}$ and $\mathbf{Y}$ have different LID values at distance r. Although the total probability measures within distance r are the same (that is, $F_{\mathbf{X}}(r) = F_{\mathbf{Y}}(r)$), $ID_{F_{\mathbf{Y}}}(r)$ is greater than one would expect for a locally uniform distribution of points in $\mathbb{R}^2$, while $ID_{F_{\mathbf{X}}}(r)$ is less.

Moreover, let $c > 1$ *be a constant, and assume that* $ID_F(0)$ *exists. Then*

$$\lim_{\substack{w \to 0^+ \\ 0 < w/c \le r \le cw}} G_{F,0,w}(r) \;=\; 1\,.$$

Proof: (Adapted from [13].)
For any $r \in (0, z)$,

$$\begin{aligned} F(r) &= F(r) \cdot \exp\left(\ln F(r) - \ln F(r) + \ln F(w) - \ln F(w)\right) \\ &\qquad \cdot \exp\left(ID_F(v)\ln w - ID_F(v)\ln w + ID_F(v)\ln r - ID_F(v)\ln r\right) \\ &= F(w) \cdot \left(\frac{r}{w}\right)^{ID_F(v)} \cdot \exp\left(ID_F(v)\ln w - ID_F(v)\ln r - \ln F(w) + \ln F(r)\right) \\ &= F(w) \cdot \left(\frac{r}{w}\right)^{ID_F(v)} \cdot \exp\left(ID_F(v)\int_r^w \frac{1}{t}\,dt - \int_r^w \frac{F'(t)}{F(t)}\,dt\right)\,, \end{aligned}$$

since F is assumed to be positive and differentiable over an open interval containing $[\min\{r, w\}, \max\{r, w\}]$. Definition 1 therefore implies that $ID_F(t)$ must exist over this interval, and thus

$$F(r) \;=\; F(w) \cdot \left(\frac{r}{w}\right)^{ID_F(v)} \cdot \exp\left(\int_r^w \frac{ID_F(v) - ID_F(t)}{t}\,dt\right)$$

as required.

For the second part of the proof, it suffices to show that

$$\int_r^w \frac{ID_F(0) - ID_F(t)}{t}\,dt \;\to\; 0\,.$$

Since $ID_F(0)$ is assumed to exist, and since F is assumed to be positive and differentiable over the open interval defined relative to 0, Definition 1 implies that ID_F must exist over the interval. Therefore, for any real value $\epsilon > 0$ there must exist a value $\delta > 0$ such that $t < \delta$ implies that $|ID_F(t) - ID_F(0)| < \epsilon$. Hence, when $cw < \delta$,

$$\left|\int_r^w \frac{ID_F(0) - ID_F(t)}{t}\,dt\right| \;\le\; \epsilon \cdot \left|\int_r^w \frac{1}{t}\,dt\right| \;=\; \epsilon \cdot \left|\ln\frac{w}{r}\right|\,.$$

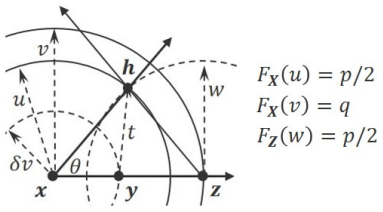


Figure 3: An illustration of the construction by which the reference point **x** is perturbed towards a target point **z**, so that the rank of **z** relative to the resulting perturbed point **y** is at most a target value.

From the limit conditions, we have that $1/c \leq w/r \leq c$, and thus $\ln(w/r)$ is bounded from above and below by constants. Therefore, since ϵ can be made arbitrarily small, the limit is indeed 0, and the result follows. □

Practical methods that have been developed for the estimation of the index, including expansion-based estimators [1] and the well-known Hill estimator and its variants [15], can all be applied to LID (for a survey, see [8]).

3 Neighborhood perturbation theorem

In this section, we present the main theoretical contribution of the paper, which provides conditions for which the perturbation of a test point can reduce the rank (by expectation) of a target location. The theorem is not directly concerned with the effect of perturbation on fixed point sets; rather, it relates to the underlying distribution from which the data can be regarded to be a sample. Our result shows that for smooth distributions over Euclidean spaces, as the intrinsic dimensionality of the test point rises, the amount of perturbation required tends to zero.

Consider a Euclidean vector space $\mathcal{S}$ with distance metric $d(\mathbf{x}, \mathbf{y}) \triangleq \|\mathbf{x} - \mathbf{y}\|$ and probability measure μ. For a given reference point $\mathbf{x} \in \mathcal{S}$ within the space, we denote by $\mathbf{X}$ the random variable associated with the distribution of distances from $\mathbf{x}$ induced by μ. The cumulative distribution function of $\mathbf{X}$ will be denoted by $F_{\mathbf{X}}$.

We begin by giving a technical lemma that establishes a condition by which a perturbation of **x** into **y** can reduce the probability of a data sample falling within a distance determined by a target location **z**.

Lemma 2 *Let* **x** *be a reference point within the Euclidean space* $\mathcal{S}$*. Let* $0 < \delta < 1/2$ *be a fixed real value, and let* p *and* q *be real values such that* $0 < p < q < 1$*. Consider the following construction with points* **y** *and* **z** *and distance parameters* r*,* u*,* v*, and* w *all depending on* p*,* q *and* δ *(see Figure 3):*

1. *Let* u *and* v *be the infimums of the distances from* **x** *at which* $F_{\mathbf{X}}(u) = p/2$ *and* $F_{\mathbf{X}}(v) = q$*, respectively.*

2. *Let* $\mathbf{z} \in \mathcal{S}$ *be any point for which* $d(\mathbf{x}, \mathbf{z}) = v$*.*

3. *Let* $\mathbf{y} \in \mathcal{S}$ *be the point lying on the line segment joining* $\mathbf{x}$ *and* $\mathbf{z}$*, at distance* $\delta\nu$ *from* $\mathbf{x}$*.*

4. *Let* r *be the supremum of the distances from* $\mathbf{y}$ *at which* $F_{\mathbf{Y}}(r) = p$*, and let* w *be the infimum of the distances from* $\mathbf{z}$ *at which* $F_{\mathbf{Z}}(w) = p/2$*.*

If $\delta \geq {}^{(\nu^2-u^2)}/_{(\nu^2-u^2+w^2)}$*, then* $d(\mathbf{y},\mathbf{z}) \leq r$*.*

Proof: Consider the two balls $B(\mathbf{x},u)$ and $B(\mathbf{z},w)$. The union of the two balls has as its axis of rotational symmetry the unique line l containing u and w. Consider any two-dimensional plane containing the line l. Within this plane, the intersection of the balls form disks as shown in Figure 3. The disk formed by the intersection of the plane with $B(\mathbf{x},u)$ will be denoted by $C(\mathbf{x},u)$; the other disks will be described in a similar manner.

The interiors of the two disks $C(\mathbf{x},u)$ and $C(\mathbf{z},w)$ must intersect. Otherwise, $u+w \leq \nu$ would hold, from which it would follow that $\nu^2 \geq (u+w)^2 \geq u^2+w^2$, and thus that $(\nu^2-u^2)/(\nu^2-u^2+w^2) \geq 1/2$. However, this would violate the assumptions on δ.

Since $B(\mathbf{x},u)$ and $B(\mathbf{z},w)$ both have probability measure $p/2$, the disk $C(\mathbf{x},u)$ cannot be contained in the strict interior of the disk $C(\mathbf{z},w)$ — otherwise, the radius of $B(\mathbf{z},w)$ could be decreased without reducing the probability associated with it, contradicting the minimality of w. $C(\mathbf{x},u)$ and $C(\mathbf{z},w)$ therefore must intersect in either two points (in which case the intersection points form two symmetric triangles with $\mathbf{x}$ and $\mathbf{z}$, one of which is illustrated in Figure 3) or one point (in which case the triangle is degenerate). Without loss of generality, let $\mathbf{h}$ be one of these intersection points.

Let t denote the distance $d(\mathbf{y},\mathbf{h})$. Regardless of whether the triangle $\triangle\mathbf{hxz}$ is degenerate, the disk $C(\mathbf{y},t)$ must be contained in the union of $C(\mathbf{x},u)$ and $C(\mathbf{z},w)$, and hence $F_{\mathbf{Y}}(t) \leq F_{\mathbf{X}}(u) + F_{\mathbf{Z}}(w) = p$. Since r is assumed to be the supremum of distances from $\mathbf{y}$ at which $F_{\mathbf{Y}}(r) = p$, we see that $r \geq t$.

Consider now the angle $\theta = \angle\mathbf{hxy} = \angle\mathbf{hxz}$, which falls in the range $0 < \theta \leq \pi$ ($\theta = \pi$ if $\triangle\mathbf{hxz}$ is degenerate). From the construction, we have

$$\begin{aligned} t^2 &= u^2 + \delta^2\nu^2 - 2\delta u\nu\cos\theta \quad \text{and} \\ w^2 &= u^2 + \nu^2 - 2u\nu\cos\theta\,, \end{aligned}$$

regardless of whether $\triangle\mathbf{hxz}$ is degenerate. Solving each equation for $2\delta u\nu\cos\theta$ and combining, we obtain

$$\begin{aligned} u^2 + \delta^2\nu^2 - t^2 &= \delta\,(u^2+\nu^2-w^2) \\ t^2 &= (1-\delta)\,u^2 + (\delta^2-\delta)\,\nu^2 + \delta w^2\,. \end{aligned} \tag{2}$$

Since $\nu > u$ implies that $\nu^2 - u^2 + w^2$ is positive, the assumption

$$\delta \geq \frac{\nu^2-u^2}{\nu^2-u^2+w^2}$$

can be restated as $\delta w^2 \geq (1-\delta)(v^2-u^2)$. Substitution of this inequality into Equation 2 yields

$$\begin{aligned} t^2 &\geq (1-\delta)\,u^2 + (\delta^2-\delta)\,v^2 + (1-\delta)(v^2-u^2) \\ &\geq (1-2\delta+\delta^2)\,v^2\,, \end{aligned}$$

and thus $t \geq (1-\delta)\,v$. However, since we have already shown that $r \geq t$, we conclude that $d(\mathbf{y},\mathbf{z}) = (1-\delta)\,v \leq t \leq r$, as required. □

For any choice of probabilities p and q such that $0 < p < q < 1$, Lemma 2 provides conditions on the proportion δ, which when satisfied guarantee that $F_{\mathbf{Y}}(d(\mathbf{y},\mathbf{z})) \leq p$ even when $F_{\mathbf{X}}(d(\mathbf{x},\mathbf{z})) = q > p$.

Under certain assumptions of the smoothness of the underlying data distribution, the construction of Lemma 2 can be used to relate the effect of perturbation on neighborhoods to the intrinsic dimensionality of the distance distribution from the perturbed point. For the following theorem, we will say that the local intrinsic dimensionality of $\mathcal{S}$ is itself continuous at $\mathbf{x} \in \mathcal{S}$ if the following conditions hold:

1. There exists a distance $\rho > 0$ for which all points $\mathbf{z} \in \mathcal{S}$ with $d(\mathbf{x},\mathbf{z}) \leq \rho$ admit a distance distribution whose cumulative distribution function $F_{\mathbf{Z}}$ is differentiable and positive within some open interval with lower bound 0.
2. $F_{\mathbf{Z}}$ converges in distribution to $F_{\mathbf{X}}$ as $\mathbf{z} \to \mathbf{x}$.
3. For each $\mathbf{z}$ satisfying the condition above, $\mathrm{ID}_{F_{\mathbf{Z}}}(0)$ exists.
4. $\lim_{\mathbf{z}\to\mathbf{x}} \mathrm{ID}_{F_{\mathbf{Z}}}(0) = \mathrm{ID}_{F_{\mathbf{X}}}(0)$.

Note that if the distributions $F_{\mathbf{Z}}$ are assumed to be absolutely continuous, and if uniform convergence is assumed in the second condition, then the third and fourth conditions would follow from the first two conditions.

Theorem 3 *Let* $\mathbf{x}$ *be a reference point within the Euclidean space* $\mathcal{S}$*, and let* $F_{\mathbf{X}}$ *be the cumulative distribution function of the distribution of distances from* $\mathbf{x}$*. Let us assume that the local intrinsic dimensionality of* $\mathcal{S}$ *is continuous at* $\mathbf{x}$*. For any real constant* $k > 1$*, consider the real-valued parameter* n *chosen such that* $n > k$*, and let* ρ_n *be the infimum of the distances for which the cumulative distribution function* $F_{\mathbf{X}}$ *achieves* k/n *— that is,* $\rho_n = \inf\{\rho \,|\, F_{\mathbf{X}}(\rho) = k/n\}$.

Let $0 < \delta < 1/2$ *be a fixed real value. With respect to the particular choice of* n*, let* $\mathbf{z}_n \in \mathcal{S}$ *be any point for which* $d(\mathbf{x},\mathbf{z}_n) = \rho_n$*, and let* $\mathbf{y}_n \in \mathcal{S}$ *be the point lying on the line segment joining* $\mathbf{x}$ *and* $\mathbf{z}_n$ *at distance* $\delta \cdot d(\mathbf{x},\mathbf{z}_n)$ *from* $\mathbf{x}$*. Then for every real value* $\varepsilon > 0$*, there exists* $n_0 > k$ *such that for all* $n \geq n_0$*, we have that*

$$\delta \geq 1 - (2k)^{\frac{-2}{\mathrm{ID}_{F_{\mathbf{X}}}(0)}} + \varepsilon \quad\Longrightarrow\quad F_{\mathbf{Y}_n}(d(\mathbf{y}_n,\mathbf{z}_n)) \leq 1/n\,.$$

Proof: For a given choice of n, consider the construction in the statement of Lemma 2, with $p = 1/n$ and $q = k/n$, where $\mathbf{z}_n = \mathbf{z}$, $\mathbf{y}_n = \mathbf{y}$, $\mathbf{h}_n = \mathbf{h}$, $v_n = d(\mathbf{x},\mathbf{z}_n) = \rho_n$, $u_n = d(\mathbf{x},\mathbf{h}_n)$, and $w_n = d(\mathbf{z}_n,\mathbf{h}_n)$. Next, we define

$$\begin{aligned} \delta_n &\triangleq (v_n^2 - u_n^2)/(v_n^2 - u_n^2 + w_n^2) \\ &= (1 - u_n^2/v_n^2)/(1 - u_n^2/v_n^2 + w_n^2/v_n^2)\,. \end{aligned} \tag{3}$$

Also, let $k_n \triangleq n \cdot F_{\mathbf{Z}_n}(\nu_n)$.

Using the local ID characterization formula of Theorem 1, we observe that

$$\begin{aligned}\frac{1}{2k} &= \frac{F_{\mathbf{X}}(u_n)}{F_{\mathbf{X}}(\nu_n)} = \left(\frac{u_n}{\nu_n}\right)^{\mathrm{ID}_{F_{\mathbf{X}}}(0)} G_{F_{\mathbf{X}},0,\nu_n}(u_n) \quad \text{and} \\ \frac{1}{2k_n} &= \frac{F_{\mathbf{Z}_n}(w_n)}{F_{\mathbf{Z}_n}(\nu_n)} = \left(\frac{w_n}{\nu_n}\right)^{\mathrm{ID}_{F_{\mathbf{Z}_n}}(0)} G_{F_{\mathbf{Z}_n},0,\nu_n}(w_n)\,,\end{aligned}$$

which in turn imply that

$$\begin{aligned}u_n/\nu_n &= (2k \cdot G_{F_{\mathbf{X}},0,\nu_n}(u_n))^{-1/\mathrm{ID}_{F_{\mathbf{X}}}(0)} \quad \text{and} \\ w_n/\nu_n &= (2k_n \cdot G_{F_{\mathbf{Z}_n},0,\nu_n}(w_n))^{-1/\mathrm{ID}_{F_{\mathbf{Z}_n}}(0)}\,.\end{aligned}$$

Substituting into Equation 3, we obtain

$$\begin{aligned}\delta_n &= \frac{1-(2k \cdot G_{F_{\mathbf{X}},0,\nu_n}(u_n))^{-2/\mathrm{ID}_{F_{\mathbf{X}}}(0)}}{1-\varepsilon_n}\,, \quad \text{where} \\ \varepsilon_n &\triangleq (2k \cdot G_{F_{\mathbf{X}},0,\nu_n}(u_n))^{-2/\mathrm{ID}_{F_{\mathbf{X}}}(0)} - \left(2k_n \cdot G_{F_{\mathbf{Z}_n},0,\nu_n}(w_n)\right)^{-2/\mathrm{ID}_{F_{\mathbf{Z}_n}}(0)}\,.\end{aligned}$$

Note that since $F_{\mathbf{Z}_n}$ is assumed to converge in distribution to $F_{\mathbf{X}}$ as $n \to \infty$,

$$\begin{aligned}\lim_{n\to\infty}\frac{k_n}{k} &= \lim_{n\to\infty}\frac{n \cdot F_{\mathbf{Z}_n}(\nu_n)}{k} \\ &= \lim_{n\to\infty}\lim_{m\to\infty}\left(\frac{F_{\mathbf{Z}_m}(\nu_n)}{F_{\mathbf{X}}(\nu_n)} \cdot \frac{F_{\mathbf{X}}(\nu_n)}{k/n}\right) \\ &= \lim_{n\to\infty}\left(\frac{F_{\mathbf{X}}(\nu_n)}{F_{\mathbf{X}}(\nu_n)} \cdot \frac{k/n}{k/n}\right) = 1\,,\end{aligned}$$

and thus $\lim_{n\to\infty} k_n = k$. Furthermore, Theorem 1 and the continuity of the intrinsic dimension of $\mathcal{S}$ imply that

$$\begin{aligned}\lim_{n\to\infty} G_{F_{\mathbf{Z}_n},0,\nu_n}(w_n) &= 1, \quad \text{and} \\ \lim_{n\to\infty} \mathrm{ID}_{F_{\mathbf{Z}_n}}(0) &= \mathrm{ID}_{F_{\mathbf{X}}}(0)\,,\end{aligned}$$

respectively. Together, these two statements establish that $\lim_{n\to\infty}\varepsilon_n = 0$, and that

$$\lim_{n\to\infty}\delta_n = 1-(2k)^{-2/\mathrm{ID}_{F_{\mathbf{X}}}(0)}\,.$$

For any real value $\varepsilon > 0$, the limit of δ_n ensures the existence of a constant $n_0 > k$ such that for all $n \geq n_0$, we have that

$$\left|\delta_n - 1 + (2k)^{-2/\mathrm{ID}_{F_{\mathbf{X}}}(0)}\right| \leq \varepsilon\,.$$

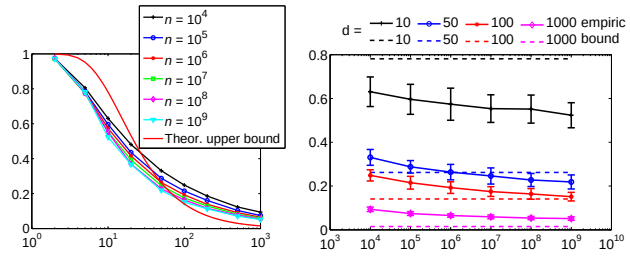
Any choice of δ satisfying

$$1/2 > \delta \geq 1-(2k)^{-2/\mathrm{ID}_{F_{\mathbf{X}}}(0)} + \varepsilon$$

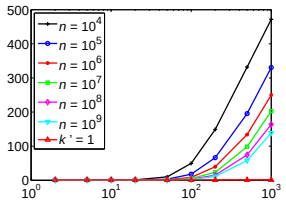
thus ensures that $\delta_n \leq \delta < 1/2$; from this, Lemma 2 can be applied with $p = 1/n$ and $q = k/n$ to yield

$$F_{\mathbf{Y}_n}(d(\mathbf{y}_n,\mathbf{z}_n)) \leq 1/n\,,$$

as required. □



(a) Mean δ (y-axis) vs. ID (x-axis) (b) Mean δ (y-axis) vs. n (x-axis)



(c) Mean k' (y-axis) vs. ID (x-axis)

Figure 4: Experiments on synthetic data.

4 Experimental validation

In this section, we design several experiments so as to verify the trends revealed by Theorem 3. This theorem should not be interpreted to mean that any given test point within a fixed data configuration always admits a perturbation that results in its k-NN object becoming its 1-NN — instead, it describes a tendency that holds asymptotically for increasingly large samples of points. Nevertheless, the theorem does illustrate an important trend: as the intrinsic dimensionality $\mathrm{ID}_{F_x}(0)$ increases, the minimum threshold on the perturbation proportion δ tends to zero.

Given a data set of size n, an embedding dimension d, and a set of n_q query points, we record the minimum perturbation proportion δ added to each query in order to reduce the rank of its k-NN to 1. Our experimental results show a clear association between the LID at the query and the amount of perturbation.

4.1 Synthetic data

We consider a simple setting involving the standardized Gaussian (normal) distribution with i.i.d. components, from which we independently draw data sets with $n \in \{10^4, 10^5, \ldots, 10^9\}$ points, and varying dimensionality $d \in \{2, 5, 10, 20, 50, 100, 200, 500, 1000\}$. The normal distribution possesses the convenient property that the local ID at each point is theoretically equal to the representational dimension d. Figure 4 shows the empirically observed trends for $n_q = 100$ query points and $k = 1000$.

Figures 4(a) and 4(b) show the observed minimum δ averaged over all query points. Figure 4(a) plots this amount against the dimensionality d for each choice of n, while

Figure 4(b) provides an alternative view of the same results by plotting (using standard deviation bars) the average minimum δ against n, for selected values of d. Two clear trends can be seen: the observed minimum δ (i) decreases with $\mathrm{ID}_{F_x}(0)$, and (ii) decreases with n. For comparison, the theoretical bound from Theorem 3 is also plotted. For low to moderate d, the observed noise levels are mostly below the theoretical bound, providing direct support to the theorem — for noise levels above the bound, the perturbation would surely produce the same effect on this data. On the other hand, although the theoretical and empirical lines cross near $d = 100$ in Figure 4(a), the second trend observed in Figure 4(b) indicates that for sufficiently-large data set sizes, the empirical noise level will eventually reach the theoretical bound.

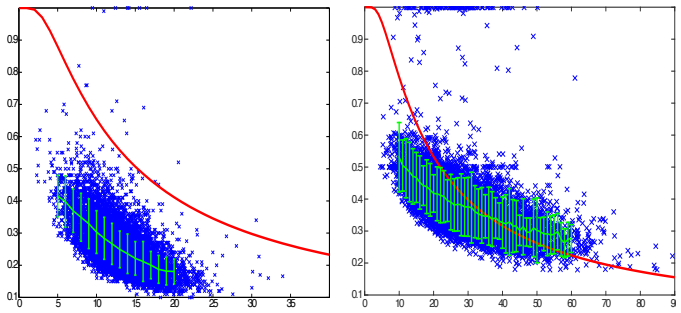
Figure 4(c) plots the average rank k' achieved by 1000-NN points after the perturbation of the query points by the amounts indicated by the theoretical bound. It can be seen that the adversarial goal of $k' = 1$ is reached for low to moderate ID, after which k' rises. However, the growth rate of k' flattens as the data set size n increases. For large ID, this trend again suggests that for sufficiently large n, perturbation by the amount given by the bound in Theorem 3 will eventually produce a rank of $k' = 1$ (by expectation). This tendency also serves to explain why the theoretical dependency between δ and ID shown in Figure 4(a) has a sharper rate of diminution than the observed dependency: for the theoretical relationship, the value of n required to achieve the perturbation goal increases with ID, whereas for the empirical relationships, n is fixed.

4.2 Real data

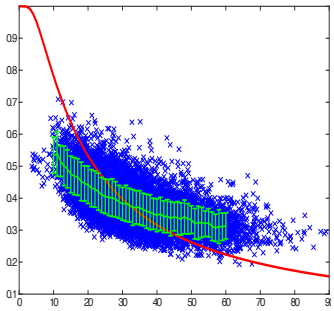
We conducted experiments with real data in order to (i) confirm the asymptotic behavior of Theorem 3 when n is extremely large, and to (ii) demonstrate that δ decreases as the local ID increases. LID values were obtained using the maximum likelihood estimator described in [1], over 100 samples.

Figure 5(a) plots the values for δ when using the BIGANN_SIFT1B dataset [16], where $d = 128$ and $n = 10^9$. Here, we chose $n_q = 10,000$ and $k = 100$. In order to estimate the mean minimum noise level, we group the ID values into integer bins. The mean and standard deviation of the noise levels for each bin is reported in this figure. In this experiment, n is extremely large, revealing the asymptotic behavior of Theorem 3. Very few values for δ are above the theoretical curve (only 22 query points). This experiment, however, uses SIFT descriptors whose estimated ID is in the low to moderate range.

In contrast, Figures 5(b) and (c) show complementary configurations where the estimated ID is much larger (although n is much smaller). Figure 5(b) plots δ against local ID for the ImageNet data set [22]. This dataset consists of $n = 1,281,167$ training images and $50,000$ test images. We take $n_q = 10,000$ images from the test set as queries. Figure 5(c) corresponds to the case of the CIFAR-10 data set [18], which consists of $n = 50,000$ training images. $10,000$ test images are also provided, which we use as queries. Both data sets were fed into a deep neural network to extract high level features. Specifically, for ImageNet data, we employed a 19-layer convolutional network [24] to extract $d = 4,096$ high level features. Similarly, for CIFAR-10, we extracted $d = 9,408$ high level features. Note that when generating adversarial high level features, the original images can be manipulated so as to conform to them, as described in [23].



(a) BIGANN_SIFT1B (b) ImageNet



(c) CIFAR-10

Figure 5: Experiments on three real data sets, plotting the minimum noise level required (y-axis) vs. estimated LID (x-axis). Red curve: theoretical bound (3). Green bars: empirical mean and standard deviation.

As expected, the theoretical curves pass through the data clouds plotted in Figures 5(b) and (c), as n is too small for the asymptotic trends to fully assert themselves. However, the plots clearly show that the amount of perturbation required for the subversion of query points decreases as the local ID grows. Note also that, as we saw in our experiments with synthetic data, the rate of diminution of the theoretical curves between δ and ID is greater than what we observed on ImageNet and CIFAR-10. Once again, this is due to the hidden dependency on the value of n needed for the theoretical relationship to hold.

5 Conclusion

In this paper, we have presented a theoretical explanation of the effect of adversarial perturbation on nearest-neighbor classification under the Euclidean distance metric: the larger the intrinsic dimensionality and data set size, the smaller the amount of adversarial noise required to transform the k-NN of a test point into a 1-NN (by expectation). These theoretical trends were confirmed experimentally for both synthetic and real data sets. Perhaps surprisingly, our result demonstrates that this vulnerability to adversarial attack

is inevitable as the data scales in both size and intrinsic dimensionality, regardless of the nature of the data.

Strictly speaking, the question remains open as to whether a quantitative explanation analogous to that of Theorem 3 can be found for other classification models, or for other similarity measures. However, it is our conjecture that the general trends should hold even for deep neural networks and other classifiers of continuously-distributed data. Intuitively, even when the distance is not Euclidean, and even when the component of the class region containing the target is not convex, an argument similar to (but perhaps considerably looser than) that of Lemma 2 is likely to hold, provided that a transformation exists between the original domain and an appropriate Euclidean domain. Theorem 3 could then be applied within the Euclidean domain, which under reverse transformation would serve to establish the trends in the original domain. The details would depend very much on the interplay between the underlying data distribution and data model, and so we will not pursue them here.

Sophisticated features, such as the ones resulting from a deep learning process, are often very effective in classification and recognition tasks. Our analysis suggests that their higher dimensionality, however, renders them very vulnerable to adversarial attack. For this reason, for deep neural networks and other state-of-the-art classifiers, a systematic and comprehensive empirical investigation of the relationship between intrinsic dimensionality and adversarial perturbation would be a very worthwhile topic for future research.

Acknowledgments

Laurent Amsaleg and Teddy Furon supported by French project Secular ANR-12-CORD-0014. James Bailey, Sarah Erfani and Nguyen Xuan Vinh supported by the Australian Research Council via grant number DP140101969. Michael E. Houle supported by JSPS Kakenhi Kiban (A) Research Grant 25240036 and Kiban (B) Research Grant 15H02753. Miloš Radovanović supported by the Serbian Ministry of Education, Science and Technological Development through project no. OI174023. Nguyen Xuan Vinh supported by a University of Melbourne ECR grant.

References

[1] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K. Kawarabayashi, and M. Nett. Estimating local intrinsic dimensionality. In *KDD*, pages 29–38, 2015.

[2] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(5):572–575, 1998.

[3] F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(10):1404–1407, 2002.

[4] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Trans. Information Theory*, 13(1):21–27, 1967.

[5] K. Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons, 2003.

[6] C. Faloutsos and I. Kamel. Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension. In *PODS*, pages 4–13, 1994.

[7] K. Fukunaga and D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *Transactions on Computers*, C-20(2):176–183, 1971.

[8] M. I. Gomes *et al*., Statistics of extremes for IID data and breakthroughs in the estimation of the extreme value index: Laurens de Haan leading contributions. *Extremes*, 11:3–34, 2008.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

[10] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1–2):189–208, 1983.

[11] A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *FOCS*, pages 534–543, 2003.

[12] M. E. Houle. Dimensionality, discriminability, density & distance distributions. In *ICDMW*, pages 468–473, 2013.

[13] M. E. Houle. Inlierness, outlierness, hubness and discriminability: an extreme-value-theoretic foundation. Technical Report NII-2015-002E, NII, Mar 2015.

[14] M. E. Houle, H. Kashima, and M. Nett. Generalized expansion dimension. In *ICDMW*, pages 587–594, 2012.

[15] R. Huisman, K. G. Koedijk, C. J. M. Kool, and F. Palm. Tail-index estimates in small samples. *Journal of Business and Economic Statistics*, 19(2):208–216, 2001.

[16] H. Jégou, R. Tavenard, M. Douze, and L. Amsaleg. Searching in one billion vectors: Re-rank with source coding. In *ICASSP*, pages 861–864, 2011.

[17] D. R. Karger and M. Ruhl. Finding nearest neighbors in growth-restricted metrics. In *STOC*, pages 741–750, 2002.

[18] A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Master's thesis, University of Toronto, 2009.

[19] D. Lowd and C. Meek. Adversarial learning. In *KDD*, pages 641–647, 2005.

[20] E. Pettis, T. Bailey, A. Jain, and R. Dubes. An intrinsic dimensionality estimator from nearest-neighbor information. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1:25–37, 1979.

[21] A. Rozza, G. Lombardi, C. Ceruti, E. Casiraghi, and P. Campadelli. Novel high intrinsic dimensionality estimators. *Machine Learning*, 89(1-2):37–65, 2012.

[22] O. Russakovsky *et al*., ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[23] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet. Adversarial manipulation of deep representations. *CoRR*, abs/1511.05122, 2015.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[25] C. J. Stone. Consistent parametric regression. *Annals of Statistics*, 5(4):595–645, 1977.

[26] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[27] P. Tabacof and E. Valle. Exploring the space of adversarial images. *CoRR*, abs/1510.05328, 2015.

[28] P. Verveer and R. Duin. An evaluation of intrinsic dimensionality estimators. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(1):81–86, 1995.