The Vulnerability of Learning to Adversarial Perturbation Increases with Intrinsic Dimensionality

Laurent Amsaleg, James Bailey, Dominique Barbe, Sarah Erfani, Michael Houle, Vinh Nguyen, Miloš Radovanovic

The Vulnerability of Learning to Adversarial Perturbation Increases with Intrinsic Dimensionality

Laurent Amsaleg*, James Bailey†, Dominique Barbe‡, Sarah Erfani†, Michael E. Houle§, Vinh Nguyen† and Miloš Radovanović¶
*CNRS-IRISA, Rennes, France
†Dept. of Computing and Information Systems, The University of Melbourne, Australia
‡ENS Rennes, France
§National Institute of Informatics, Tokyo, Japan
¶Dept. of Mathematics and Informatics, University of Novi Sad, Serbia

***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. It has also been observed that these adversarial samples generalize well across models. A complete understanding of the nature of adversarial samples has not yet emerged. Towards this goal, we present a novel theoretical result formally linking the adversarial vulnerability of learning to the intrinsic dimensionality of the data. In particular, our investigation 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.**

I. Introduction

Recent research has shown that machine learning systems, including state-of-the-art deep neural networks, can be subverted when a small amount of carefully-designed, imperceptible adversarial noise is added to an input object so as to influence a classification result [1]. Adversarial perturbation generalizes surprisingly well across different models [2]. These alarming observations have many practical implications in an era where machine learning technologies are ubiquitous.

It has been attempted to explain adversarial perturbation from different perspectives. Because adversarial samples tend to generalize well across models [2], explanations involving overfitting and/or the peculiarity of individual learning systems have been dismissed. Moreover, it has recently been shown that even models with parameters picked at random are unstable with respect to adversarial perturbation [3]. In [4], 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 [4]. 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.

This paper presents, 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 dimensionality (ID) [5], [6]. In the context of Euclidean spaces, this paper provides a constructive proof 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 paper is organized as follows. In Section II, we give a brief overview of adversarial perturbation and the concept of intrinsic dimensionality, as well as some of the useful properties of the LID model. In Section III, we give our main theoretical result, followed in Section IV by an experimental validation of the impact of intrinsic dimensionality on the adversarial perturbation effect. Section V concludes with a discussion of some of the possible implications of our result for deep neural networks and other state-of-the-art learning systems.

II. Background

A. Adversarial perturbation

For a general machine learning model, many adversarial perturbation strategies are possible, such as the one presented here. Following the notation in [7], 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 $\mathbf{p}^A = [\mathbb{1}_{i=a}]$, with zero probabilities to all but a chosen adversarial label a. One way to craft the adversarial noise $\mathbf{d}$ is by solving the following optimization problem:

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

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 [2].

In this paper, we propose 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 [8], [9]. 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 necessarily the candidate closest to the test point), and move the test point toward the target point along the straight line joining them. 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, 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.

B. Intrinsic dimensionality

Over the past decades, many characterizations of the ID of sets have been proposed [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. Projection-based learning methods such as PCA [16] can produce as a byproduct an estimate of ID. Expansion-based models include the expansion dimension (ED) [20], the generalized expansion dimension (GED) [21], and the minimum neighbor distance (MiND) [18].

As a motivating example from m-dimensional Euclidean space, consider the situation in 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 as the numbers of points they enclose [21], [20].

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 [6], [5]. The notion of volume is naturally analogous to that of probability measure. ID can then be modeled as a function of distances $r > 0$, 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 ([5]): Let F be the cumulative distribution function of a random distance variable. If there exists an open interval I containing $r > 0$ over which F is non-zero and continuously differentiable, then the *local intrinsic dimensionality (LID)* of F at r is given by

$$\mathrm{ID}_F(r) \triangleq \lim_{\epsilon \to 0^+} \frac{\ln\left(F((1+\epsilon)r)/F(r)\right)}{\ln(1+\epsilon)} = \frac{r \cdot F'(r)}{F(r)}.$$

The second equality follows by applying l'Hôpital's rule to the limit.

Under this assumption of distributional smoothness (continuous differentiability), the original data set determines a sample of distances from a given point, for which the intrinsic dimensionality (here referred to simply as 'local ID', or 'LID') of the cumulative distribution F is estimated. The definition of ID_F can be extended to the case where $r = 0$ by taking the limit of $\mathrm{ID}_F(r)$ as $r \to 0^+$, whenever this limit exists:

$$\mathrm{ID}_F^* \triangleq \lim_{r \to 0^+} \mathrm{ID}_F(r)\,.$$

Figure 1 illustrates the local ID of distance distributions.

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 [22], it is shown that the EVT representation of the cumulative distribution F completely determines function ID_F, and that the EVT index is in fact identical to ID_F^*.

Theorem 1 ([22]): Let $F : \mathbb{R}^{\ge 0} \to \mathbb{R}$ be a real-valued function such that ID_F^* exists. Let r and w be positive values for which $F(r)$ and $F(w)$ are both positive. If F is non-zero and continuously differentiable everywhere in an open interval containing $[\min\{r,w\}, \max\{r,w\}]$, then

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

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

Fig. 1: 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, $F_{\mathbf{x}}(r) = F_{\mathbf{y}}(r)$), $\mathrm{ID}_{F_{\mathbf{Y}}}(r)$ is greater than one would expect for a locally uniform distribution of points in $\mathbb{R}^2$, while $\mathrm{ID}_{F_{\mathbf{X}}}(r)$ is less.

whenever the integral exists.

Moreover, let $c > 1$ be a constant. Then

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

Practical methods that have been developed for the estimation of the EVT index, including expansion-based estimators [6] and the well-known Hill estimator and its variants [23], can all be applied to LID (for a survey, see [24]).

III. 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 alter the rank (by expectation) of a target location with respect to this test point. The theorem to be presented 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 results show that for smooth distributions over Euclidean spaces, as the intrinsic dimensionality at the test point rises, the amount of perturbation required tends to zero. These distributions are assumed to be smooth in two senses at once:

1) The distributions of distances from the target location and test location (both before and after perturbation) must have cumulative distribution functions that satisfy the LID smoothness assumptions.
2) The LID values must themselves be continuous over some open interval containing the original test point. The precise notion of continuity will be introduced in Section III-B.

It should be noted that unlike classical treatments of intrinsic dimensionality in machine learning in which the data is assumed to be restricted to a Riemannian (smooth) manifold of a given (intrinsic) dimensionality, our distributional assumptions are in fact much more general.

A. Perturbation and distribution

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 $F_{\mathbf{x}}$ the cumulative distribution function associated with the distribution of distances from $\mathbf{x}$, as induced by μ.

We begin by giving one technical lemma that establishes conditions by which a perturbation of $\mathbf{x}$ into a new point $\mathbf{y} \in \mathcal{S}$ can affect the expected ranking relationships with respect to a target location $\mathbf{z} \in \mathcal{S}$. The expected rank of an item is taken to be the number of samples m times the probability of a sample falling closer to the reference point than the target point. More precisely, if $F_{\mathbf{x}}$ is the cumulative distribution function associated with the distribution of distances from $\mathbf{x}$, then the expected rank of $\mathbf{z}$ relative to $\mathbf{x}$ is $m \cdot F_{\mathbf{x}}(\|\mathbf{z} - \mathbf{x}\|)$. Since we are reasoning about distributions and not samples, for convenience we will treat m as a fixed but unknown quantity, and refer to $F_{\mathbf{x}}(\|\mathbf{z} - \mathbf{x}\|)$ as the *distributional rank* of $\mathbf{z}$ relative to $\mathbf{x}$.

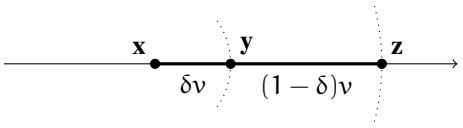


Fig. 2: Relationships among the reference point $\mathbf{x}$, its perturbation $\mathbf{y}$, and the target location $\mathbf{z}$.

To carry on with this discussion we now need to define:

- p and q to be probability values such that $0 < p < q < 1$;
- $\nu > 0$ to be a distance value;
- $0 < \delta < 1$ to be a fixed real proportion of the distance ν.

We begin by choosing a point $\mathbf{z}$ at distance ν from $\mathbf{x}$ such that the ball of radius ν centered at $\mathbf{x}$ has probability measure equal to q. We then define certain perturbations of $\mathbf{x}$ that produce a new point $\mathbf{y}$, always at distance $\delta\nu$ from $\mathbf{x}$. Finally, we give sufficient conditions on δ such that the distance $\|\mathbf{z} - \mathbf{y}\|$ satisfies a certain locality criterion relative to $\mathbf{y}$, involving the probability p. Figure 2 shows the different relationships among $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, ν, and δ.

The following lemma shows that a sufficiently-large perturbation of the test point $\mathbf{x}$ directly toward a target location $\mathbf{z}$ can decrease the distributional rank of the target from the probability q (relative to the test point) to less than the probability p (relative to the perturbed point $\mathbf{y}$).

Lemma 2: Consider the following construction depending on $\mathbf{x}$, p, q and δ:

1) Let ν be a distance from $\mathbf{x}$ at which $F_{\mathbf{x}}(\nu) = q$.
2) Let $\mathbf{z} \in \mathcal{S}$ be any point for which $\|\mathbf{z} - \mathbf{x}\| = \nu$.
3) Let $\mathbf{y} \in \mathcal{S}$ be the point at distance $\delta\nu$ from $\mathbf{x}$ lying in the segment joining $\mathbf{x}$ and $\mathbf{z}$.
4) Let r be the infimum of the distances from $\mathbf{y}$ at which $F_{\mathbf{y}}(r) = p$.

If $\delta > 1 - r/\nu$, then $F_{\mathbf{y}}(\|\mathbf{z} - \mathbf{y}\|) < p$.

Proof: This situation is illustrated in Figure 2.

Since $F_{\mathbf{y}}(r') = p$ is satisfied for $r' = r$, but for no distance values $r' < r$, the monotonicity of the cumulative distribution function $F_{\mathbf{y}}$ ensures that

$$\|\mathbf{z}-\mathbf{y}\| < r \iff F_{\mathbf{y}}(\|\mathbf{z}-\mathbf{y}\|) < p\,.$$

From the assumptions on $\mathbf{y}$, $\mathbf{z}$, v, δ, and r, and from the collinearity of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, we obtain

$$\begin{aligned} F_{\mathbf{y}}(\|\mathbf{z}-\mathbf{y}\|) < p &\iff \|\mathbf{z}-\mathbf{x}\| - \|\mathbf{y}-\mathbf{x}\| < r \\ &\iff v(1-\delta) < r\,. \end{aligned}$$

We thus conclude that if $\delta > 1 - r/v$, then $F_{\mathbf{y}}(\|\mathbf{z}-\mathbf{y}\|) < p$, as required. □

B. Asymptotic effects of perturbation

We will say that the local intrinsic dimensionality 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 $\|\mathbf{z}-\mathbf{x}\| \leq \rho$ admit a distance distribution whose cumulative distribution function $F_{\mathbf{z}}$ is continuously 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, $\mathrm{ID}^*_{F_{\mathbf{z}}}$ exists.
4) $\lim_{\mathbf{z}\to\mathbf{x}} \mathrm{ID}^*_{F_{\mathbf{z}}} = \mathrm{ID}^*_{F_{\mathbf{x}}}$.

We denote by $F_{\mathbf{s}}$ the cumulative distribution function of the distribution of distances induced at $\mathbf{s} \in \mathcal{S}$. We also assume the existence of some reference point $\mathbf{x} \in \mathcal{S}$ for which the local intrinsic dimensionality is continuous.

It is possible to use Lemma 2 to show that as the number of data samples increases, a sufficiently-large perturbation of the test point directly towards a target location of distributional rank k/n (or equivalently, expected neighbor rank k) will eventually reduce the distributional rank of the target to below $1/n$ (expected neighbor rank 1).

Theorem 3: Given a real constant $k > 1$, let n be a real-valued parameter chosen such that $n > k$, and let v_n be a distance for which the cumulative distribution function $F_{\mathbf{x}}$ achieves the value k/n.

Let $\delta > 0$ 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 $\|\mathbf{z}_n - \mathbf{x}\| = v_n$, and let $\mathbf{y}_n \in \mathcal{S}$ be a point at distance δv_n from $\mathbf{x}$ lying on the segment joining $\mathbf{x}$ and $\mathbf{z}_n$. Then for every real value $\varepsilon > 0$, there exists $n_0 > k$ such that for all $n \geq n_0$, we have

$$\delta > 1 - k^{\frac{-1}{\mathrm{ID}^*_{F_{\mathbf{x}}}}} + \varepsilon \implies F_{\mathbf{y}_n}(\|\mathbf{z}_n - \mathbf{y}_n\|) < 1/n\,.$$

When n tends toward infinity, then the minimum value for δ as to satisfy the condition in the statement of the Theorem 3 is $1 - k^{-1/\mathrm{ID}^*_{F_{\mathbf{x}}}}$.

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 $\mathbf{z}_n = \mathbf{z}$, $\mathbf{y}_n = \mathbf{y}$, $v_n = \|\mathbf{z}_n - \mathbf{x}\| = v$, and $r_n = r$ (as illustrated in Figure 2 with $\mathbf{y}_n = \mathbf{y}$ and $\mathbf{z}_n = \mathbf{z}$). Note that this construction implies that $F_{\mathbf{x}}(v_n) = k/n$ and $F_{\mathbf{y}_n}(r_n) = 1/n$. In addition, we define $\alpha_n \triangleq n \cdot F_{\mathbf{y}_n}(v_n)$.

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

$$\frac{1}{\alpha_n} = \frac{F_{\mathbf{y}_n}(r_n)}{F_{\mathbf{y}_n}(v_n)} = \left(\frac{r_n}{v_n}\right)^{\mathrm{ID}^*_{F_{\mathbf{y}_n}}} \cdot G_{F_{\mathbf{y}_n},0,v_n}(r_n)\,;$$

rearranging, we obtain

$$\frac{r_n}{v_n} = \left(\frac{1}{\alpha_n \cdot G_{F_{\mathbf{y}_n},0,v_n}(r_n)}\right)^{1/\mathrm{ID}^*_{F_{\mathbf{y}_n}}}\,.$$

Since $F_{\mathbf{y}_n}$ is assumed to converge in distribution to $F_{\mathbf{x}}$ as $n \to \infty$, we have that

$$\begin{aligned} \lim_{n\to\infty} \alpha_n &= \lim_{n\to\infty} n \cdot F_{\mathbf{y}_n}(v_n) \\ &= \lim_{n\to\infty} \lim_{m\to\infty} \left(\frac{F_{\mathbf{y}_m}(v_n)}{F_{\mathbf{x}}(v_n)} \cdot \frac{F_{\mathbf{x}}(v_n)}{1/n}\right) \\ &= \lim_{n\to\infty} \left(\frac{F_{\mathbf{x}}(v_n)}{F_{\mathbf{x}}(v_n)} \cdot \frac{k/n}{1/n}\right) \\ &= k\,. \end{aligned}$$

Furthermore, Theorem 1 and the continuity of the intrinsic dimension of $\mathcal{S}$ imply that

$$\begin{aligned} \lim_{n\to\infty} G_{F_{\mathbf{y}_n},0,v_n}(r_n) &= 1\,, \quad \text{and} \\ \lim_{n\to\infty} \mathrm{ID}^*_{F_{\mathbf{y}_n}} &= \mathrm{ID}^*_{F_{\mathbf{x}}}\,. \end{aligned}$$

Defining $\delta_n \triangleq 1 - r_n/v_n$, these two statements establish that

$$\lim_{n\to\infty} \delta_n = 1 - k^{-1/\mathrm{ID}^*_{F_{\mathbf{x}}}}\,.$$

For any real value $\varepsilon > 0$, the limit of δ_n ensures the existence of a constant $n_0 > n_1$ such that for all $n \geq n_0$, we have that

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

Any choice of δ satisfying

$$\delta > 1 - k^{-1/\mathrm{ID}^*_{F_{\mathbf{x}}}} + \varepsilon$$

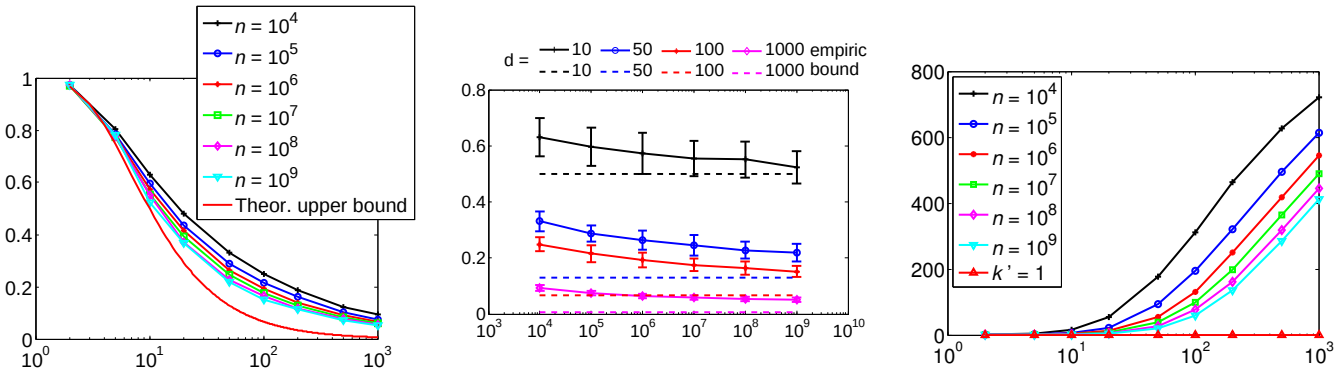
thus ensures that $\delta > \delta_n$; from this, Lemma 2 can be applied with $p = 1/n$ and $q = k/n$ to yield

$$F_{\mathbf{y}_n}(\|\mathbf{z}_n - \mathbf{y}_n\|) < 1/n$$

as required. □

We have seen that in terms of distributional rank, the amount of perturbation required to transform a 1-nearest-neighbor location into a k-nearest-neighbor location depends on the local ID of the test point; moreover, as the local ID and number of sample points both tend to infinity, the amount of perturbation required diminishes to zero.

We observe that instead of considering perturbations that involve distributional ranks of $1/n$ and k/n, we can instead use distributional ranks of any p and q such that $0 < p < q < 1$. The bounds stated in the theorem would be the same, except that every occurrence of k would be replaced by q/p.



(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)

Fig. 3: Experiments on synthetic data.

IV. 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 illustrates an important trend: as the intrinsic dimensionality $\mathrm{ID}^*_{F_x}$ 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.

A. 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. In Figure 3 we show the empirically observed trends for $n_q = 100$ query points and $k = 1000$.

Figures 3(a) and 3(b) show the observed minimum δ averaged over all query points. Figure 3(a) plots this amount against the dimensionality d for each choice of n, while Figure 3(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 $\mathrm{ID}^*_{F_x}$, and (ii) decreases with n. For comparison, the theoretical bound for δ is also plotted. For all values of d, the observed noise levels are in each case above the theoretical bound, providing direct support to the theorem.

Figure 3(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 3(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.

B. 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 [6].

Figure 4 plots the values for δ when using the BIGANN_SIFT1B dataset [25], 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 below the theoretical curve; however, we note that it is always possible for some values to be below the curve, due to such influences as the use of observed neighbor rank as an estimate of distributional rank, and error in the estimation of LID values.

V. Conclusion

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). This theoretical trend is confirmed experimentally for both synthetic and real data sets.

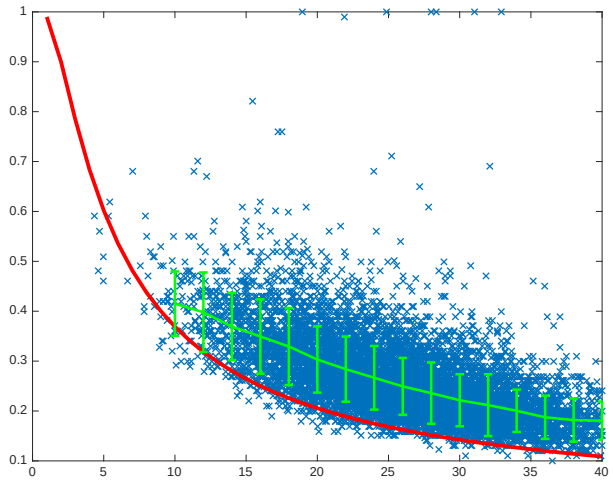


Fig. 4: Experiments on the BIGANN_SIFT1B real data sets, plotting the minimum noise level required (y-axis) vs. estimated LID (x-axis). Red curve: theoretical bound from Theorem 3. Green bars: empirical mean and standard deviation.

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 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 is in part supported by the European CHIST-ERA ID_IOT project. James Bailey, Sarah Erfani and Vinh Nguyen are in part supported by the Australian Research Council via grant number DP140101969. Vinh Nguyen is in part supported by a University of Melbourne ECR grant. Michael E. Houle is in part supported by JSPS Kakenhi Kiban (A) Research Grant 25240036 and Kiban (B) Research Grant 15H02753. Miloš Radovanović is in part supported by the Serbian Ministry of Education, Science and Technological Development through project number OI174023.

References

[1] D. Lowd and C. Meek, "Adversarial learning," in *KDD*, 2005.
[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *CoRR*, vol. abs/1312.6199, 2013.
[3] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet, "Adversarial manipulation of deep representations," *CoRR*, vol. abs/1511.05122, 2015.
[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2014.
[5] M. E. Houle, "Dimensionality, discriminability, density & distance distributions," in *ICDMW*, 2013.
[6] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K. Kawarabayashi, and M. Nett, "Estimating local intrinsic dimensionality," in *KDD*, 2015.
[7] P. Tabacof and E. Valle, "Exploring the space of adversarial images," *CoRR*, vol. abs/1510.05328, 2015.
[8] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE TIT*, vol. 13, no. 1, 1967.
[9] C. J. Stone, "Consistent parametric regression," *Annals of Statistics*, vol. 5, no. 4, 1977.
[10] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons, 2003.
[11] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," *Physica D: Nonlinear Phenomena*, vol. 9, no. 1–2, 1983.
[12] F. Camastra and A. Vinciarelli, "Estimating the intrinsic dimension of data with a fractal-based method," *IEEE TPAMI*, vol. 24, no. 10, 2002.
[13] C. Faloutsos and I. Kamel, "Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension," in *PODS*, 1994.
[14] A. Gupta, R. Krauthgamer, and J. R. Lee, "Bounded geometries, fractals, and low-distortion embeddings," in *FOCS*, 2003.
[15] J. Bruske and G. Sommer, "Intrinsic dimensionality estimation with optimally topology preserving maps," *IEEE TPAMI*, vol. 20, no. 5, 1998.
[16] K. Fukunaga and D. R. Olsen, "An algorithm for finding intrinsic dimensionality of data," *IEEE TOC*, vol. C-20, no. 2, 1971.
[17] E. Pettis, T. Bailey, A. Jain, and R. Dubes, "An intrinsic dimensionality estimator from nearest-neighbor information," *IEEE TPAMI*, vol. 1, 1979.
[18] A. Rozza, G. Lombardi, C. Ceruti, E. Casiraghi, and P. Campadelli, "Novel high intrinsic dimensionality estimators," *Machine Learning*, vol. 89, no. 1-2, 2012.
[19] P. Verveer and R. Duin, "An evaluation of intrinsic dimensionality estimators," *IEEE TPAMI*, vol. 17, no. 1, 1995.
[20] D. R. Karger and M. Ruhl, "Finding nearest neighbors in growth-restricted metrics," in *STOC*, 2002.
[21] M. E. Houle, H. Kashima, and M. Nett, "Generalized expansion dimension," in *ICDMW*, 2012.
[22] M. E. Houle, "Local intrinsic dimensionality I: An extreme-value-theoretic formulation for similarity applications," in *SISAP*, 2017.
[23] R. Huisman, K. G. Koedijk, C. J. M. Kool, and F. Palm, "Tail-index estimates in small samples," *Journal of Business and Economic Statistics*, vol. 19, no. 2, 2001.
[24] M. I. Gomes, L. Canto e Castro, M. I. Fraga Alves, and D. Pestana, "Statistics of extremes for IID data and breakthroughs in the estimation of the extreme value index: Laurens de Haan leading contributions," *Extremes*, vol. 11, 2008.
[25] H. Jégou, R. Tavenard, M. Douze, and L. Amsaleg, "Searching in one billion vectors: Re-rank with source coding," in *ICASSP*, 2011.