High Intrinsic Dimensionality Facilitates Adversarial Attack: Theoretical Evidence

Laurent Amsaleg, James Bailey, Amelie Barbe, Sarah Erfani, Teddy Furon, Michael Houle, Milos Radovanovic, Nguyen Xuan Vinh

▶ **To cite this version:**

Laurent Amsaleg, James Bailey, Amelie Barbe, Sarah Erfani, Teddy Furon, et al.. High Intrinsic Dimensionality Facilitates Adversarial Attack: Theoretical Evidence. IEEE Transactions on Information Forensics and Security, 2020, 16, pp.1-12. 10.1109/TIFS.2020.3023274 . hal-02938099v2

HAL Id: hal-02938099
https://hal.science/hal-02938099v2

Submitted on 8 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

High Intrinsic Dimensionality Facilitates Adversarial Attack: Theoretical Evidence

Laurent Amsaleg, *Member, IEEE*, James Bailey, *Member, IEEE*, Amélie Barbe, Sarah M. Erfani, Teddy Furon, *Senior Member, IEEE*, Michael E. Houle, *Member, IEEE*, Miloš Radovanović, and Xuan Vinh Nguyen

***Abstract*— Machine learning systems are vulnerable to adversarial attack. By applying to the input object a small, carefully-designed perturbation, a classifier can be tricked into making an incorrect prediction. This phenomenon has drawn wide interest, with many attempts made to explain it. However, a complete understanding is yet to emerge. In this paper we adopt a slightly different perspective, still relevant to classification. We consider retrieval, where the output is a set of objects most similar to a user-supplied query object, corresponding to the set of k-nearest neighbors. We investigate the effect of adversarial perturbation on the ranking of objects with respect to a query. Through theoretical analysis, supported by experiments, we demonstrate that as the intrinsic dimensionality of the data domain rises, the amount of perturbation required to subvert neighborhood rankings diminishes, and the vulnerability to adversarial attack rises. We examine two modes of perturbation of the query: either 'closer' to the target point, or 'farther' from it. We also consider two perspectives: 'query-centric', examining the effect of perturbation on the query's own neighborhood ranking, and 'target-centric', considering the ranking of the query point in the target's neighborhood set. All four cases correspond to practical scenarios involving classification and retrieval.**

***Index Terms*— Adversarial attack, intrinsic dimensionality, nearest neighbor.**

I. Introduction

RECENT research has shown that the performance of machine learning systems, including state-of-the-art deep neural networks (DNNs), can be subverted by a form of adversarial attack. Such an attack adds a small, carefully-designed, adversarial perturbation to an input object, so as to influence a classification result [1], [2]. In the case of images, the perturbation is imperceptible to humans and can deceive a high-performing image classifier into misclassifying a test image to any other desired class. This phenomenon has become a major security concern for real-world applications of DNNs, such as self-driving cars and identity recognition [3], [4].

Attempts have been made to explain adversarial attack from different perspectives. It has been demonstrated that almost all machine learning approaches are sensitive to adversarial attack [5], [6], including linear classifiers, SVMs, decision trees, and deep neural networks. It has also been demonstrated that adversarial samples generalize well across classifiers, architectures, training sets, and even machine learning approaches based on different paradigms [2], [7]–[10].

Deep neural networks are well-known for their excellent classification performance. Consequently, many theoretical contributions as well as more empirical investigations, have focused on understanding adversarial perturbation in the context of deep learning (DL). DL architectures have been deconstructed, and most of their elements, functions, model parameters, assumptions and mechanisms scrutinized as possible origins of the sensitivity of DL to adversarial attacks [11]. Complicating the analysis, it has been found that even models with parameters picked at random are unstable with respect to adversarial perturbations [12].

The search for an explanation of adversarial perturbation has also investigated characterizing the decision boundaries between classes. The curvature of class boundaries, in particular, seems to contribute significantly to the adversarial effect [13]–[15]. Perhaps more importantly, 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 [7]. Reference [16] highlight that adversarial subspaces lie near (but not on) the data sub-manifold. A theoretical study of adversarial examples was conducted on a synthetic data set [17] representing two concentric m-dimensional spheres. It was proved that any model that erroneously classifies a small fraction of the data manifold, is also vulnerable to adversarial perturbations of size $O(1/\sqrt{m})$. Adversarial perturbation in the context of a k-NN classifier was analyzed in [18], which found that robustness characteristics are closely tied to the

The work of Laurent Amsaleg was supported in part by the European CHIST-ERA ID_IOT project. The work of James Bailey and Sarah M. Erfani was supported in part by the Australian Research Council under Grant DP140101969. The work of Teddy Furon was supported by the ANR-AID Chaire SAIDA. The work of Michael E. Houle was supported by the JSPS Kakenhi Kiban (B) Research under Grant 18H03296. The work of Miloš Radovanović was supported by the Serbian National Project under Grant OI174023.

Laurent Amsaleg and Teddy Furon are with Inria, CNRS, IRISA, Campus de Beaulieu, Univ Rennes, 35042 Rennes, France (e-mail: laurent.amsaleg@irisa.fr; teddy.furon@irisa.fr).

James Bailey and Sarah M. Erfani are with the School of Computing and Information Systems, The University of Melbourne, Parkville, VIC 3010, Australia (e-mail: baileyj@unimelb.edu.au; sarah.erfani@unimelb.edu.au).

Amélie Barbe is with the Laboratoire de Physique, École Normale Supérieure de Lyon, 69364 Lyon, France (e-mail: amelie.barbe@ens-lyon.fr).

Michael E. Houle is with the National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: meh@nii.ac.jp).

Miloš Radovanović is with the Faculty of Sciences, University of Novi Sad, 21000 Novi Sad, Serbia (e-mail: radacha@dmi.uns.ac.rs).

Xuan Vinh Nguyen is with NVIDIA Corporation, Santa Clara, CA 95051 USA (e-mail: vinhn@nvidia.com).

Digital Object Identifier 10.1109/TIFS.2020.3023274

chosen value of k, and that there exist different regimes of behavior — one regime for when k is constant, the other for when $k = \Omega(\sqrt{mn \log n})$, where m is the data dimension and n is the sample size. Adversarial resistance has also been investigated in the context of deep k-NN networks [19], and gradient descent based attacks have recently been developed that target this type of network [20].

Despite the hypotheses and insights so far proposed in the literature, gaining a full understanding of the causes and behavior of the adversarial perturbation effect remains a challenging and important open problem.

A. From Classification to Retrieval

The studies discussed above are concerned with *classification* systems, and the effects of erroneous class label assignments. This paper adopts a slightly different point of view, still relevant to the classification setting: it is concerned with *retrieval*, where the output of the system is not a class label, but a set of objects most similar to a user-supplied query object, according to some predefined measure of (dis)similarity. This is traditionally exemplified by the set of k-nearest neighbors (k-NN) of the query point, often (but not necessarily) according to Euclidean distance. Retrieval within high-dimensional data domains is essential to algorithms that are at the core of many data analysis procedures, including indexing, (subspace) clustering, outlier detection, and data mining, as well as classification.

Retrieval is also central to sensitive applications, such as the enforcement of the protection of copyrights, in which the task is to detect close copies of objects stored in a database. Once detected, the upload of copyrighted material can be blocked, or alternatively, monetized. All content identification applications, such as those involving biometry of faces or irises, use some form of similarity retrieval for filtering illegal content.

The field of adversarial retrieval addresses several forms of attack: the hiding of data from queries ('evasion'), the seeding of training sets with false data ('poisoning'), and the falsification of classification decisions (either targeted at specific classes, or indiscriminate). Adversarial attack perturbs the location of points, which impacts neighborhood set memberships and rankings, which in turn undermines the ability to determine the similarity of objects. Very early works have investigated issues relating to attack on content-based retrieval systems [21]–[23]. Other forms of attack have targeted face detection and recognition using camouflage art, also known as computer vision dazzle [24]–[26].

This paper investigates the effect of adversarial perturbation on the ranking of objects with respect to a query, such as may arise in both machine learning classification contexts and more general retrieval. Through a theoretical analysis, we show that as the intrinsic dimensionality of the data domain rises, the amount of perturbation required to subvert neighborhood rankings diminishes, and thus the vulnerability of queries to adversarial attack tends to rise. We examine two modes of perturbation of the query: either 'closer' to the target point, or 'farther' from it. We also consider two vantage points for the effect: 'query-centric,' in which we examine the effect of perturbation on the query's own neighborhood ranking, and 'target-centric', in which we consider the ranking of the query point in the target's neighborhood set. This gives a total of four cases to be addressed: *closer + query*, *closer + target*, *farther + query*, and *farther + target*.

The aforementioned cases for our analysis all have relevance in real-world applications of content-based retrieval and classification. Adversarial perturbation of a query point in order for it to be misclassified by a nearest-neighbor classifier is a simple example of the *closer + query* scenario. In content-based retrieval, a small modification of a query image away from a copyrighted target photo might prevent the identification of the query as a quasi-copy of the target, thereby bypassing filters or monetization mechanisms (*farther + query*). Conversely, perturbations of images can be performed before insertion into a database so that image would become a 'hub' object [27] appearing in an excessively-large proportion of retrieval results (*closer + target*). Such images could be created by dishonest content owners wishing to increase revenue through promotion. Another possibility is to inject adversarial content into databases perturbed so as to appear in as few query neighborhoods as possible (a so-called 'anti-hub' [27]). If the perturbation is successful, the content would be concealed from all but the most precise queries — in this way, hiding illicit content from the general public would be possible, whereas informed users could still access it (*farther + target*).

B. Contributions

In a preliminary version of this paper, in the context of Euclidean spaces, we provided a theoretical explanation of the adversarial effect of perturbation for the *closer + query* scenario [28], in terms of the Local Intrinsic Dimensionality (LID) [29]–[31]. The LID characterizes the order of magnitude of the growth of probability measure with respect to a neighborhood of increasing radius. In this journal version, besides improving the previous result, we extend the analysis to all four of the scenarios described above. This paper and its antecedent present new theoretical explanations of the adversarial effect in similarity-based classification and retrieval.

The analysis deals with distributions of inter-point distances and not fixed point sets per se. The notion of neighbor is considered in terms of mathematical expectation, as follows: with respect to a sample size n, a target location $\mathbf{z}$ is a k-nearest neighbor (k-NN) of reference point $\mathbf{x}$ *by expectation* if k out of n sample points would be expected to lie within distance $d(\mathbf{x}, \mathbf{z})$ of $\mathbf{x}$.

Constructive proofs are provided within which methods are given for perturbing a reference point $\mathbf{x}$ to location $\mathbf{y}$, so that by expectation as $n \to \infty$:

- the k-NN of query $\mathbf{x}$ becomes the 1-NN of the perturbed query $\mathbf{y}$ (the *closer + query* scenario);
- $\mathbf{x}$ is the k-NN of a targeted location, but $\mathbf{y}$ becomes its 1-NN (*closer + target*);
- the 1-NN of query $\mathbf{x}$ becomes the k-NN of the perturbed query $\mathbf{y}$ (*farther + query*);

- **x** is the 1-NN of a targeted location, but **y** becomes its k-NN (*farther + target*).

Conditions on **y** are provided for a relationship to hold between the required amount of perturbation on the one hand, and on the other, the LID of the distance distribution from **z** or from **y**. Our analysis leads to the conclusion that *as the intrinsic dimensionality rises, the relative amount of perturbation required to achieve the desired effect tends to zero*.

C. Differences With Previous Studies

Our analysis differs with the previous theoretical studies focusing on adversarial classification in large dimensional spaces. In [6], Fawzi *et al.* analyze the robustness to perturbation of classifiers on data produced according to a generative model, whose latent distribution is assumed to be Gaussian. Their analysis makes use of the Gaussian isoperimetric inequality. The analysis of Shafahi *et al.* [32] resembles that of [6] in that it too is based on isoperimetric inequalities, except that the data points are distributed over the hypersphere. Their setups do not hold asymptotically. In [17], Gilmer *et al.* consider only a very specific data point distribution, the 'concentric spheres dataset'. The setup holds asymptotically with respect to the representational dimension. In the context of metric probability spaces, Mahloujifar *et al.* [33] blend the aforementioned approaches, by arguing that the isoperimetric inequalities hold approximately over a much larger distribution family as the dimension m goes to infinity.

Our work differs from the aforementioned results, in that:

- it does not rely on a particular choice of data point distribution (such as Gaussians or other ideal distributions, or uniformity over the hypersphere or concentric spheres);
- it therefore is not based on any statistical concentration phenomenon;
- it holds asymptotically with respect to the dataset size n, as the dimension m of the space is fixed.

D. Structure of the Article

The remainder of the paper is organized as follows. Section II gives a brief overview of feature vector extraction and the concept of local intrinsic dimensionality, together with a review of some of the useful properties of the LID model. In Section III we present the statements and proofs of our main theoretical results. This is followed in Section IV by an experimental validation of the impact of intrinsic dimensionality on the adversarial perturbation effect. Section V concludes the paper with a discussion of some of the possible implications of our results for deep neural networks and other state-of-the-art learning systems, as well as adversarial retrieval.

II. Background

A. Feature Vector Extraction

The core function of a general content retrieval architecture is that of feature extraction, in which a global vector representation is generated from content according to its nature (such as text, audio, or still images) and the definition of content similarity (such as in topic, in style, or in color). Ideally, similar contents should be represented by feature vectors close together in the feature space, and dissimilar contents by vectors far from each other. In this way, the task of similar content retrieval ultimately reduces to that of feature vector retrieval according to an appropriate feature vector extraction function. We denote the extraction function by $f : \mathcal{C} \to \mathbb{R}^m$, which produces a vector $\mathbf{x}$ of m feature values from a content object $\mathbf{c} \in \mathcal{C}$.

The lemmas and theorems presented in the following sections deal with points in Euclidean space, which we interpret as the underlying feature representations of objects in some content-based classification or retrieval context.

B. Intrinsic Dimensionality

Over the past decades, many characterizations have been proposed for the global intrinsic dimensionality of sets or the local intrinsic dimensionality of a point in a set (see [34] and the references therein). This section summarizes the generalized expansion dimension (GED) [34] and its extension to continuous distance distributions, the local intrinsic dimensionality (LID) [29], [30].

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:

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

For a finite data set, GED formulations are obtained by estimating the volume of balls by the numbers of data set points they enclose [34], [35].

Instead of regarding intrinsic dimensionality as a characteristic of a collection of data points (as evidenced by their distances) the GED was recently extended to a statistical setting, in which the distribution of distances from a given point is modeled as a continuous random variable [29], [36]. 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 $\epsilon \to 0^+$. For an illustration of the intrinsic dimensionality of distance distributions, see Figure 1.

Definition 1 ([29], [30]): Let F be a real-valued function. 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.

In this paper, we will make use of the local intrinsic dimensionality model to characterize the properties of a distribution in the vicinity of a point of interest within its domain, in terms of a secondary or 'local' distribution of distances induced by the original, 'global' distribution.

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.

Let $\mathcal{D}$ be a distribution over a domain $\mathcal{S}$ equipped with distance measure $d(\mathbf{x}, \mathbf{y})$. With respect to a given reference point $\mathbf{x} \in \mathcal{S}$, $\mathcal{D}$ induces a univariate distribution $\mathcal{D}_{\mathbf{x}}$ taking values determined by distances relative to $\mathbf{x}$. More precisely, the event of drawing sample $\mathbf{y} \in \mathbb{R}^m$ from the 'global' distribution $\mathcal{D}$ determines the event of generating the value $d(\mathbf{x}, \mathbf{y})$ in the 'local' distribution $\mathcal{D}_{\mathbf{x}}$. For a given point $\mathbf{x} \in \mathcal{S}$ within the domain, we denote by $F_{\mathbf{x}}$ the cumulative distribution function of $\mathcal{D}_{\mathbf{x}}$; given a value r, $F_{\mathbf{x}}(r)$ is the probability that a sample $\mathbf{y}$ drawn from the 'global' distribution $\mathcal{D}$ lies within distance r of $\mathbf{x}$ — that is, the probability of satisfying $d(\mathbf{x}, \mathbf{y}) \leq r$.

The local intrinsic dimensionality with respect to $\mathbf{x}$, denoted as $\mathrm{LID}(\mathbf{x})$, characterizes the close neighborhood of point $\mathbf{x}$ by taking the limit of $\mathrm{ID}_{F_{\mathbf{x}}}(r)$ as $r \to 0^+$, whenever this limit exists:

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

Although for the purposes of this paper we will assume that the domain is the vector space $\mathbb{R}^m$ with the Euclidean norm $d(\mathbf{x}, \mathbf{y}) \triangleq \|\mathbf{x} - \mathbf{y}\|$, the LID model is in fact more general, and places no assumptions on the distribution $\mathcal{D}_{\mathbf{x}}$ or distance measure $d(\mathbf{x}, \mathbf{y})$ beyond what is necessary for the definition of $\mathrm{LID}(\mathbf{x})$ to hold.

The smallest distances from point $\mathbf{x}$ 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 [30], it is shown that the EVT representation of the cumulative distribution $F_{\mathbf{x}}$ completely determines function $\mathrm{ID}_{F\mathbf{x}}$, and that the EVT index is in fact identical to $\mathrm{LID}(\mathbf{x})$.

Theorem 1 [30]: Let $F : \mathbb{R}^{\geq 0} \to \mathbb{R}$ be a real-valued function, and assume that $\mathrm{ID}_F(0)$ 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

$$\frac{F(r)}{F(w)} = \left(\frac{r}{w}\right)^{\mathrm{ID}_F(0)} \cdot G_{F,w}(r), \text{ where}$$
$$G_{F,w}(r) \triangleq \exp\left(\int_r^w \frac{\mathrm{ID}_F(0) - \mathrm{ID}_F(t)}{t}\,\mathrm{d}t\right) ,$$

whenever the integral exists.

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

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

Given a data sample $S \subseteq \mathcal{S}$ and a reference point $\mathbf{x} \in \mathcal{S}$, the distributional interpretation sees the observed distances from $\mathbf{x}$ to the vectors of S as samples of the distribution $F_{\mathbf{x}}$, from which the intrinsic dimensionality is estimated. Practical methods that have been developed for the estimation of the EVT index, including expansion-based estimators [36] and the well-known Hill estimator and its variants [37], can all be applied to LID (for a survey, see [38]). Recently, techniques have been developed that use expansion from neighboring points to stabilize LID estimation, allowing for smaller neighborhood samples to be used [39].

III. Neighborhood Perturbation Theorems

A. Distance and Distributional Rank

This section presents the main theoretical contribution of the paper. It provides conditions for which the perturbation of a test point can affect its 'rank' within a distribution of distances from some reference location $\mathbf{z} \in \mathcal{S}$. With respect to $\mathbf{z} \in \mathcal{S}$, any point $\mathbf{x} \in \mathcal{S}$ determines a distance $r = \|\mathbf{x} - \mathbf{z}\|$, which in turn determines a probability $p = F_{\mathbf{z}}(r)$. The point $\mathbf{x}$ (or more precisely, its distance r to $\mathbf{z}$) determines at the p-th quantile of the distance distribution $\mathcal{D}_{\mathbf{z}}$: compared to $\mathbf{x}$, any sample drawn from the global distribution $\mathcal{D}$ has probability p of being closer to $\mathbf{z}$. Therefore, with regard to the distribution $\mathcal{D}_{\mathbf{z}}$, the point $\mathbf{x}$ is given a '*distributional rank*' within the space of all samples in $\mathcal{S}$, in which the proportion of samples ahead of it in the ranking is simply the value p determining its quantile within $\mathcal{D}_{\mathbf{z}}$.

We begin our analysis with the effect of perturbation on the distributional rank (or quantile), rather than on the traditional distance-based neighbor ranking determined with respect to some fixed sample set. However, for a given point $\mathbf{x}$ determining the p-th distance quantile (that is, $r = \|\mathbf{x} - \mathbf{z}\|$ such that $p = F_{\mathbf{z}}(r)$), the probability p does determine the *expected rank* of $\mathbf{x}$, which is naturally defined as the expected number of sample points having distance to $\mathbf{z}$ that is smaller than that of $\mathbf{z}$ (which for a sample of size n is simply np). It is for this notion of expected rank that we ultimately derive our main results. Note that the question of quantile estimation, rank variance,

TABLE I

SUMMARY OF THE NOTATION USED IN SECTIONS II AND III

$\mathrm{ID}_F(r)$	The local intrinsic dimensionality of function F at r.
$\mathrm{ID}_F(0)$	The local intrinsic dimensionality of function F, defined as $\lim_{r\to 0^+}\mathrm{ID}_F(r)$.
I	An interval of distance values.
$\mathcal{S}$	A domain.
d	A distance measure associated with domain $\mathcal{S}$.
$\mathbf{x}$	A point within domain $\mathcal{S}$.
$\mathcal{D}$	A distribution within domain $\mathcal{S}$.
$\mathcal{D}_{\mathbf{x}}$	The distribution of distances from $\mathbf{x}\in\mathcal{S}$ induced by $\mathcal{D}$.
$F_{\mathbf{x}}$	The cumulative distribution function of $\mathcal{D}_{\mathbf{x}}$.
$\mathrm{LID}(\mathbf{x})$	$\mathrm{ID}_{F_{\mathbf{x}}}(0)$, the local intrinsic dimensionality with respect to $\mathbf{x}$.
$G_{F,w}(r)$	An auxiliary function defined in the statement of Theorem 1.
S	A sample of points drawn from distribution $\mathcal{D}$.
n	The size of sample S.
r	A generic value, which in some contexts is interpreted as a distance.
p	A generic probability value.
$\mathbf{z}$	A point within domain $\mathcal{S}$ in the vicinity of $\mathbf{x}$ and $\mathbf{y}$.
$\mathbf{y}$	A point within $\mathcal{S}$ obtained through the perturbation of $\mathbf{x}$.
v	The distance between $\mathbf{x}$ and $\mathbf{z}$.
δ	A proportion, such that the distance between $\mathbf{x}$ and $\mathbf{y}$ is δv.
ψ	The angle based at $\mathbf{x}$ formed by the rays to $\mathbf{y}$ and $\mathbf{z}$.
p_x	A probability value, interpreted as a distributional rank involving $\mathbf{z}$ and $\mathbf{x}$ (before perturbation).
p_y	A probability value, interpreted as a distributional rank involving $\mathbf{z}$ and $\mathbf{y}$ (after perturbation).
p_t	A probability value, interpreted as a target distributional rank involving $\mathbf{z}$ (after perturbation).
$[\underline{r},\overline{r}]$	The interval of distances at which the cumulative distribution function $F_{\mathbf{z}}$ achieves a target probability p_t.
ρ	The radius of a small neighborhood centered at $\mathbf{x}$.
ε	A small positive distance value less than ρ.
k_x	For a sample of size n, the expected rank np_x associated with the distributional rank p_x.
k_y	For a sample of size n, the expected rank np_y associated with the distributional rank p_y.
k_t	For a sample of size n, the expected rank np_t associated with the target distributional rank p_t.
n_0	A minimum sample size above which the claims of Theorem 2 and Corollary 3 are shown to hold.
w	A small positive bound on the difference between a local intrinsic dimensionality at distance u, and its limit as $u\to 0$.
φ	The minimum of two ratios of distances, both >1.

and other distributional properties of fixed finite samples has been well studied elsewhere (see for example [40]).

For two different perturbation strategies, the results presented in this section indicate that as the local intrinsic dimensionality rises, the amount of perturbation required to significantly modify an expected rank tends to zero. For the analysis to hold, the distance distributions are assumed to be smooth in two senses at once:

1) The distributions of distances must have cumulative distribution functions that satisfy the LID smoothness assumptions (*cf.* Def. 1 and Th. 1)
2) The LID values must themselves be continuous over some open interval containing the original point. The precise notion of continuity will be introduced in Section III-C.

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 local distributional assumptions are in fact much more general.

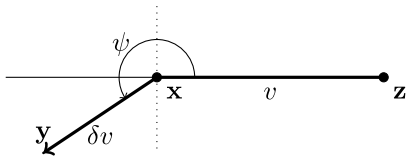


Fig. 2. Relationships among the point $\mathbf{x}$, its perturbed version $\mathbf{y}$, and the location $\mathbf{z}$.

B. Perturbation and Distribution

We begin by establishing a technical lemma about the conditions by which a perturbation moving $\mathbf{x}$ onto $\mathbf{y}$ affects the expected ranking relationships between $\mathbf{y}$ and location $\mathbf{z}\in\mathcal{S}$.

Lemma 1 considers a point $\mathbf{x}$ at distance v from $\mathbf{z}$. A perturbation of $\mathbf{x}$ produces a new point $\mathbf{y}$, whose distance from $\mathbf{x}$ can be expressed as δv for some proportion $\delta>0$. The lemma gives sufficient conditions on δ to substantially change the distributional rank. Figure 2 illustrates the different relationships among $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, v, and δ assumed in the statement of the lemma.

The lemma shows that a sufficiently-large perturbation pushing $\mathbf{x}$ to a point $\mathbf{y}$ *closer to* (resp. *farther from*) $\mathbf{z}$ decreases (resp. increases) the distributional rank relative to $F_{\mathbf{z}}$ from $F_{\mathbf{z}}(\|\mathbf{x}-\mathbf{z}\|)$ to at most $p_t < F_{\mathbf{z}}(\|\mathbf{x}-\mathbf{z}\|)$ (resp. at least $p_t > F_{\mathbf{z}}(\|\mathbf{x}-\mathbf{z}\|)$).

Lemma 1: Consider the following construction depending on $\mathbf{x}$, $\mathbf{z}$ and $\delta>0$:

1) Let $p_x = F_{\mathbf{z}}(\|\mathbf{x}-\mathbf{z}\|)$, and let $v=\|\mathbf{x}-\mathbf{z}\|$.
2) Let $\mathbf{y}\in\mathcal{S}$ be any point at distance δv from $\mathbf{x}$.
3) Let $p_t\in(0,1)$ be the targeted distributional rank of $\mathbf{y}$ relative to $\mathbf{z}$.
4) Let $I=[\underline{r},\overline{r}]$ be the interval of the distances from $\mathbf{z}$ for which $F_{\mathbf{z}}(r)=p_t, \forall r\in I$.

Then, the construction satisfies

- for $p_t\le p_x$ (*closer to*): $F_{\mathbf{z}}(\|\mathbf{y}-\mathbf{z}\|) < p_t$ whenever $\delta > 1-\underline{r}/v$.
- for $p_t\ge p_x$ (*farther from*): $F_{\mathbf{z}}(\|\mathbf{y}-\mathbf{z}\|) > p_t$ whenever $\delta > \overline{r}/v-1$.

Proof: Denote ψ the angle between $\mathbf{y}-\mathbf{x}$ and $\mathbf{z}-\mathbf{x}$ as illustrated in Figure 2. By construction, we obtain:

$$\begin{aligned}\|\mathbf{y}-\mathbf{z}\|^2 &= \|\mathbf{y}-\mathbf{x}+\mathbf{x}-\mathbf{z}\|^2\\ &= \|\mathbf{y}-\mathbf{x}\|^2+\|\mathbf{z}-\mathbf{x}\|^2-2\,(\mathbf{y}-\mathbf{x})\cdot(\mathbf{z}-\mathbf{x})\\ &= \delta^2v^2+v^2-2\delta v^2\cos\psi\ .\end{aligned}$$

In the first case, we push $\mathbf{y}$ towards $\mathbf{z}$ so that its distributional rank is lower than the target p_t. The monotonicity of the cumulative distribution function $F_{\mathbf{z}}$ and the definition of the interval I, ensures that:

$$\begin{aligned}F_{\mathbf{z}}(\|\mathbf{y}-\mathbf{z}\|)<p_t &\iff \|\mathbf{y}-\mathbf{z}\|<\underline{r}\\ &\iff \delta^2-2\delta\cos\psi+\left(1-\underline{r}^2/v^2\right)<0.\end{aligned}$$

Trivially, if the starting point $\mathbf{x}$ already has a distributional rank $p_x<p_t$, it implies that $v<\underline{r}$ as well, and the above inequality

is satisfied for $\delta = 0$. More interestingly, when $p_x \geq p_t$, the inequality is satisfied for the following conditions:

$$|\sin\psi| < \underline{r}/\upsilon$$
$$|\delta - \cos\psi| < \sqrt{\underline{r}^2/\upsilon^2 - \sin^2\psi}\,.$$

Not surprisingly, the lower bound on δ is minimized for $\psi = 0$ — that is, when $\mathbf{y}$ is pushed directly towards $\mathbf{z}$. This allows us to state that, for any admissible angle ψ, we have $\delta > 1 - \underline{r}/\upsilon$. Note that the upper bound on δ is not meaningful if larger than 1. Pushing $\mathbf{y}$ exactly onto $\mathbf{z}$ makes $\delta = 1$, and $F_{\mathbf{z}}(\|\mathbf{y} - \mathbf{z}\|) = 0$ — that is, always lower than $p_t > 0$.

In the second case, we push $\mathbf{y}$ away from $\mathbf{z}$ so that its distributional rank is larger than the target p_t. The monotonicity of the cumulative distribution function $F_{\mathbf{z}}$ and the definition of the interval I, ensures that:

$$\begin{aligned} F_{\mathbf{z}}(\|\mathbf{y} - \mathbf{z}\|) > p_t &\iff \|\mathbf{y} - \mathbf{z}\| > \overline{r} \\ &\iff \delta^2 - 2\delta\cos\psi + \left(1 - \overline{r}^2/\upsilon^2\right) > 0. \end{aligned}$$

Trivially, if the starting point $\mathbf{x}$ already has a distributional rank $p_x > p_t$, it implies that $\upsilon > \overline{r}$ as well, and the above inequality is satisfied for $\delta = 0$. More interestingly, when $p_x \leq p_t$ so that $\overline{r}/\upsilon \geq 1$, the polynomial $\delta^2 - 2\delta\cos\psi + \left(1 - \overline{r}^2/\upsilon^2\right)$ has two real roots, and is strictly positive if and only if δ is strictly outside their interval. The lowest root being negative, this means that δ is greater than the upper root:

$$\delta > \cos\psi + \sqrt{\overline{r}^2/\upsilon^2 - \sin^2\psi}.$$

Not surprisingly, this lower bound is minimized for $\psi = \pi$, *i.e.* $\mathbf{x}$ is pushed in the opposite direction of $\mathbf{z}$. This allows us to state that, for any angle ψ, $\delta > \overline{r}/\upsilon - 1$. ∎

This lemma considers the changes after perturbation from $\mathbf{x}$ to $\mathbf{y}$ in the distributional rank relative to a nearby fixed vantage point $\mathbf{z}$. A variant of the lemma can be stated that instead considers the change in distributional rank of $\mathbf{z}$ when the vantage point shifts from $\mathbf{x}$ to $\mathbf{y}$. We omit the proof, since the techniques are entirely analogous to those of Lemma 1.

Lemma 2: Consider the following construction depending on $\mathbf{x}$, $\mathbf{z}$ and $\delta > 0$:

1) Let $p_x = F_{\mathbf{x}}(\|\mathbf{z} - \mathbf{x}\|)$, and let $\upsilon = \|\mathbf{z} - \mathbf{x}\|$.
2) Let $\mathbf{y} \in \mathcal{S}$ be any point at distance $\delta\upsilon$ from $\mathbf{x}$.
3) Let $p_t \in (0, 1)$ be the targeted distributional rank of $\mathbf{z}$ relative to $\mathbf{y}$.
4) Let $I = [\underline{r}, \overline{r}]$ be the interval of the distances from $\mathbf{y}$ for which $F_{\mathbf{y}}(r) = p_t, \forall r \in I$.

Then, the construction satisfies

- for $p_t \leq p_x$ (*closer to*): $F_{\mathbf{y}}(\|\mathbf{z} - \mathbf{y}\|) < p_t$ whenever $\delta > 1 - \underline{r}/\upsilon$.
- for $p_t \geq p_x$ (*farther from*): $F_{\mathbf{y}}(\|\mathbf{z} - \mathbf{y}\|) > p_t$ whenever $\delta > \overline{r}/\upsilon - 1$.

C. Effects of Perturbation on Expected Rank

We now turn our attention to the relationship between local intrinsic dimensionality and the effect of perturbations on neighborhoods, under certain assumptions of the smoothness of the underlying data distribution. We 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 $\mathbf{z} \in \mathbb{R}^m$ with $\|\mathbf{z} - \mathbf{x}\| \leq \rho$ admit a distance distribution $\mathcal{D}_{\mathbf{z}}$ whose cumulative distribution function $F_{\mathbf{z}}$ is continuously differentiable and positive within some open interval with lower bound 0.
2) For any sequence $\mathbf{z} \to \mathbf{x}$ of points satisfying Condition 1, convergence in distribution of the sequence of random distance variables defined at $\mathbf{z}$ to the distance variable defined at $\mathbf{x}$; that is, the condition $\lim_{\mathbf{z}\to\mathbf{x}} F_{\mathbf{z}}(\varepsilon) = F_{\mathbf{x}}(\varepsilon)$ holds for any distance $\varepsilon \in (0, \rho)$.
3) For each $\mathbf{z}$ satisfying Condition 1, $\mathrm{LID}(\mathbf{z})$ exists and is positive.
4) $\lim_{\mathbf{z}\to\mathbf{x}} \mathrm{LID}(\mathbf{z}) = \mathrm{LID}(\mathbf{x})$.

Condition 1 implies that $F_{\mathbf{z}}^{-1}(p_t)$ is well-defined for $p_t < F_{\mathbf{z}}(\rho)$, and that the intervals I defined in Lemmas 1 and 2 reduce to a singleton: $\underline{r} = \overline{r} = r < \rho$. For the remainder of this section, we assume that the local intrinsic dimensionality is continuous at point $\mathbf{x} \in \mathcal{S}$.

The main theorem of this paper is a theoretical statement concerning the effect of perturbation on expected rank. As in the statement of Lemma 1, we consider the situation in which a given point $\mathbf{x}$ is perturbed to a new location $\mathbf{y}$, all from the perspective of the distance distribution $\mathcal{D}_{\mathbf{z}}$ associated with a point of interest $\mathbf{z} \in \mathcal{S}$, where $\mathbf{z} \neq \mathbf{x}, \mathbf{y}$. We again denote the distances of $\mathbf{x}$ and $\mathbf{y}$ to $\mathbf{z}$ by $\upsilon = \|\mathbf{x} - \mathbf{z}\|$ and $\delta\upsilon = \|\mathbf{y} - \mathbf{z}\|$, respectively, for the appropriate choice of $\delta > 0$. In order to determine the expected ranks of $\mathbf{x}$ and $\mathbf{y}$ relative to the distribution of distances to $\mathbf{z}$, the analysis also takes into account an additional parameter: n, a sample size. Denoting the expected rank of $\mathbf{x}$ by $k_x = np_x$ (where p_x is the distributional rank of $\mathbf{x}$ with respect to $\mathcal{D}_{\mathbf{z}}$), the question tackled by the analysis is: how large must δ be to achieve a target expected rank k_t after perturbation to $\mathbf{y}$? Note that since the theorem concerns *expected rank* and not *sample rank*, the argument makes no recourse to any particular sample set.

The theorem uses Lemma 1 to show that as the number n of samples increases, a sufficiently-large perturbation of a test point closer to (respectively, away from) a target location of expected rank k_x reduces (respectively, increases) the expected rank of the target below (respectively, above) k_t. The proportional amount of perturbation required, δ, is a decreasing function of both LID and n.

Theorem 2: Let k_x, k_t, and δ be positive real constants. For any sufficiently small real value $\varepsilon > 0$, there exists a positive integer $n_0 > \max\{k_x, k_t\}$ for which the following construction holds for all choices of integer $n \geq n_0$.

Let $\mathbf{z} \in \mathcal{S}$ be any point that would achieve a fixed expected rank $k_x = n \cdot F_{\mathbf{z}}(\|\mathbf{x} - \mathbf{z}\|)$ relative to its own distance distribution $\mathcal{D}_{\mathbf{z}}$, against a sample of size n. Let $\upsilon = \|\mathbf{x} - \mathbf{z}\|$. Let $\mathbf{y} \in \mathcal{S}$ be the result of perturbing $\mathbf{x}$ by a distance $\delta\upsilon$. Then the following implications hold regarding the expected rank $k_y = n \cdot F_{\mathbf{z}}(\|\mathbf{y} - \mathbf{z}\|)$ of the perturbed point $\mathbf{y}$, in relation to a targeted expected rank $k_t \neq k_x$:

- When $k_t < k_x$ (the target expected rank is less than that of $\mathbf{x}$),

$$\delta > 1 - (k_t/k_x)^{1/\mathrm{LID}(\mathbf{x})} + \varepsilon \implies k_y < k_t\,.$$

- When $k_t > k_x$ (the target expected rank is more than that of $\mathbf{x}$),

$$\delta > (k_t/k_x)^{1/\mathrm{LID}(\mathbf{x})} - 1 + \varepsilon \implies k_y > k_t \,.$$

Proof: For a given choice of n, consider the construction in the statement of Lemma 1, with $p_x = k_x/n$ and $p_t = k_t/n$, where $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, υ and δ are as defined above, and where r is defined such that $F_{\mathbf{z}}(r) = p_t$. Using the local ID characterization formula of Theorem 1, we observe that

$$\frac{k_t}{k_x} = \frac{p_t}{p_x} = \frac{F_{\mathbf{z}}(r)}{F_{\mathbf{z}}(\upsilon)} = \left(\frac{r}{\upsilon}\right)^{\mathrm{LID}(\mathbf{z})} \cdot G_{F_{\mathbf{z}},\upsilon}(r) \,.$$

Rearranging, we obtain

$$\frac{r}{\upsilon} = \left(\frac{k_t}{k_x \cdot G_{F_{\mathbf{z}},\upsilon}(r)}\right)^{1/\mathrm{LID}(\mathbf{z})} .$$

Applying a logarithmic transformation, and substituting for $G_{F_{\mathbf{z}},\upsilon}(r)$, we arrive at

$$\ln \frac{r}{\upsilon} = \frac{1}{\mathrm{LID}(\mathbf{z})} \ln \frac{k_t}{k_x} - \frac{1}{\mathrm{LID}(\mathbf{z})} \int_r^{\upsilon} \frac{\mathrm{LID}(\mathbf{z}) - \mathrm{ID}_{F_{\mathbf{z}}}(u)}{u} \,\mathrm{d}u \,.$$

Note again that $\mathrm{LID}(\mathbf{z})$ is assumed to be positive, and also that $k_t \neq k_x$ implies that $r \neq \upsilon$. Hence, for any supplied value of ε such that $0 < \varepsilon < \frac{r}{\upsilon}$, there exists a sufficiently small distance value $w > 0$ such that for all $0 < u < w$, the absolute difference between $\mathrm{ID}_{F_{\mathbf{z}}}(u)$ and its limit $\mathrm{LID}(\mathbf{z})$ can be bounded by

$$|\mathrm{ID}_{F_{\mathbf{z}}}(u) - \mathrm{LID}(\mathbf{z})| \leq w \leq \frac{\mathrm{LID}(\mathbf{z}) \ln(\varphi)}{|\ln(\frac{\upsilon}{r})|} \,,$$

where the upper bound is a function of ε via the new real-valued variable

$$\varphi = \min \left\{ \frac{r + \upsilon\varepsilon}{r}, \frac{r}{r - \upsilon\varepsilon} \right\} > 1 \,.$$

Moreover, by choosing n_0 sufficiently large, we can guarantee that, for all $n \geq n_0$, w also satisfies

$$w \geq \max\{F_{\mathbf{z}}^{-1}(k_x/n), F_{\mathbf{z}}^{-1}(k_t/n)\} = \max\{\upsilon, r\}.$$

Therefore, we may conclude that for any choice of $\varepsilon \in (0, r/\upsilon)$, there exists $n_0 > 0$ such that, whenever $n >= n_0$,

$$\begin{aligned} \left|\ln \frac{r}{\upsilon} - \frac{1}{\mathrm{LID}(\mathbf{z})} \ln \frac{k_t}{k_x}\right| &\leq \frac{1}{\mathrm{LID}(\mathbf{z})} \left| \int_r^{\upsilon} \frac{\mathrm{LID}(\mathbf{z}) - \mathrm{ID}_{F_{\mathbf{z}}}(u)}{u} \,\mathrm{d}u \right| \\ &\leq \frac{\ln \varphi}{|\ln(\frac{\upsilon}{r})|} \left| \int_r^{\upsilon} \frac{1}{u} \,\mathrm{d}u \right| = \ln \varphi. \end{aligned}$$

Rearranging, the condition yields

$$\begin{aligned} \ln \frac{r}{\upsilon\varphi} \leq \frac{1}{\mathrm{LID}(\mathbf{z})} \ln \frac{k_t}{k_x} &\leq \ln \frac{r\varphi}{\upsilon} \\ \ln\left(\frac{r}{\upsilon} \cdot \frac{r - \upsilon\varepsilon}{r}\right) \leq \frac{1}{\mathrm{LID}(\mathbf{z})} \ln \frac{k_t}{k_x} &\leq \ln\left(\frac{r}{\upsilon} \cdot \frac{r + \upsilon\varepsilon}{r}\right) \\ \ln\left(\frac{r}{\upsilon} - \varepsilon\right) \leq \frac{1}{\mathrm{LID}(\mathbf{z})} \ln \frac{k_t}{k_x} &\leq \ln\left(\frac{r}{\upsilon} + \varepsilon\right) \\ \frac{r}{\upsilon} - \varepsilon \leq \left(\frac{k_t}{k_x}\right)^{1/\mathrm{LID}(\mathbf{z})} &\leq \frac{r}{\upsilon} + \varepsilon \,. \end{aligned}$$

Using these bounds together with Lemma 1, for $k_t < k_x$ we see that

$$\begin{aligned} \delta > 1 - (k_t/k_x)^{1/\mathrm{LID}(\mathbf{x})} + \varepsilon &\implies \delta > 1 - r/\upsilon \\ &\implies k_y < k_t \,, \end{aligned}$$

and for $k_t > k_x$ we see that

$$\begin{aligned} \delta > (k_t/k_x)^{1/\mathrm{LID}(\mathbf{x})} - 1 + \varepsilon &\implies \delta > r/\upsilon - 1 \\ &\implies k_y > k_t \end{aligned}$$

as required. ∎

Some remarks are in order:

- For a fixed choice of the ratio k_t/k_x, Theorem 2 shows that the proportion of perturbation required to achieve the target rank is a decreasing function of LID.
- For large LID, through the use of the expansion $e^u = 1 + u + o(u)$, the amount of perturbation required can be seen to scale as

$$\delta > \frac{1}{\mathrm{LID}(\mathbf{x})} \left|\ln \frac{k_t}{k_x}\right| + \varepsilon + o\left(\frac{1}{\mathrm{LID}(\mathbf{x})}\right).$$

- Using the formulation of Lemma 2 in which the roles of $F_{\mathbf{z}}$ and $F_{\mathbf{y}}$ are interchanged, together with the assumption of the local continuity of LID, the statements in Theorem 2 regarding the ranks achieved with respect to $\mathcal{D}_{\mathbf{z}}$ can be shown to apply to the ranks with respect to $\mathcal{D}_{\mathbf{y}}$.

Corollary 3: Let k_x, k_t, and δ be positive real constants. For any sufficiently small real value $\varepsilon > 0$, there exists a positive integer $n_0 > \max\{k_x, k_t\}$ for which the following construction holds for all choices of integer $n \geq n_0$.

Let $\mathbf{z} \in \mathcal{S}$ be any point that would achieve a fixed expected rank $k_x = n \cdot F_{\mathbf{z}}(\|\mathbf{x} - \mathbf{z}\|)$ relative to the distance distribution $\mathcal{D}_{\mathbf{x}}$, against a sample of size n. Let $\upsilon = \|\mathbf{z} - \mathbf{x}\|$. Let $\mathbf{y} \in \mathcal{S}$ be the result of perturbing $\mathbf{x}$ by a distance $\delta\upsilon$. Then the following implications hold regarding the expected rank $k_y = n \cdot F_{\mathbf{y}}(\|\mathbf{z} - \mathbf{y}\|)$ of relative to the distance distribution $\mathcal{D}_{\mathbf{y}}$ of the perturbed point $\mathbf{y}$, in relation to a targeted expected rank $k_t \neq k_x$:

- When $k_t < k_x$ (the target expected rank is less than that of $\mathbf{x}$),

$$\delta > 1 - (k_t/k_x)^{1/\mathrm{LID}(\mathbf{x})} + \varepsilon \implies k_y < k_t \,.$$

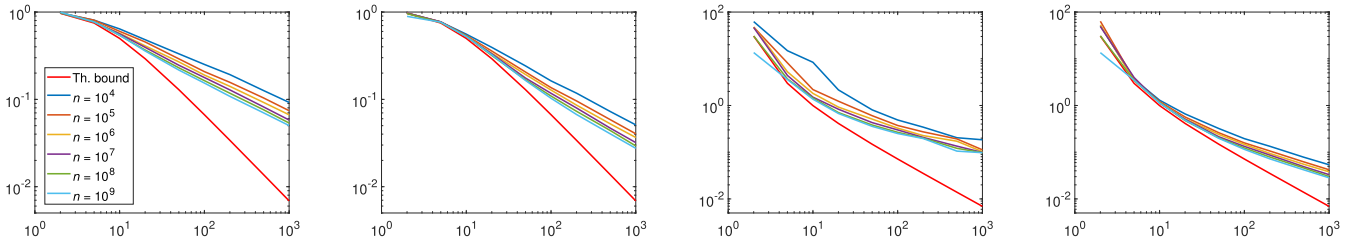
- When $k_t > k_x$ (the target expected rank is more than that of $\mathbf{x}$),

$$\delta > (k_t/k_x)^{1/\mathrm{LID}(\mathbf{x})} - 1 + \varepsilon \implies k_y > k_t \,.$$

We will later discuss some of the implications of our theory, in Section V. Next though, as a validation of the theory presented here, we empirically evaluate the effects of perturbation on real and synthetic data.

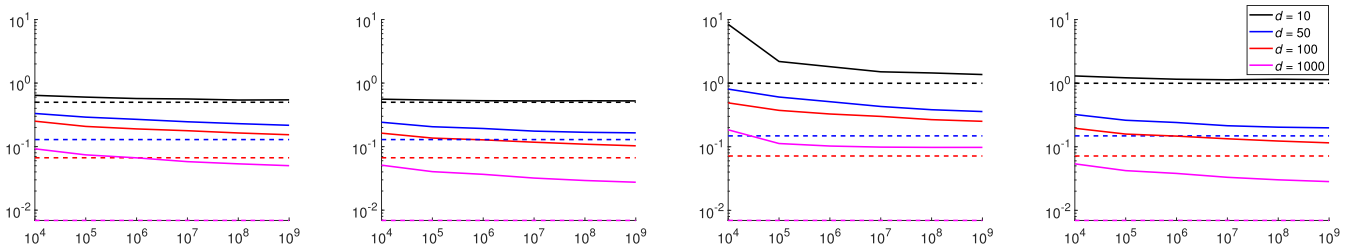
IV. Experimental Validation

Theorem 2 provides *sufficiency conditions* on the proportional perturbation δ that hold *asymptotically*, as the number of data samples tends to infinity. This theorem should therefore not be interpreted as guaranteeing the success or failure of individual perturbations in a practical scenario — they do not imply that any given test point within a fixed data



(a) *Closer+query* w.r.t. $F_{\mathbf{y}_n}$. (b) *Closer+target* w.r.t. $F_{\mathbf{z}_n}$. (c) *Farther+query* w.r.t. $F_{\mathbf{y}_n}$. (d) *Farther+target* w.r.t. $F_{\mathbf{z}_n}$.

Fig. 3. Mean perturbation δ (y-axis) vs. estimated LID (x-axis / log scale). Colors correspond to different dataset sizes. The red curve shows the theoretical sufficiency bound.



(a) *Closer+query* w.r.t. $F_{\mathbf{y}_n}$. (b) *Closer+target* w.r.t. $F_{\mathbf{z}_n}$. (c) *Farther+query* w.r.t. $F_{\mathbf{y}_n}$. (d) *Farther+target* w.r.t. $F_{\mathbf{z}_n}$.

Fig. 4. Mean perturbation δ (y-axis) vs. $\log_{10} n$ (x-axis). Colors correspond to different dataset dimensionalities. Dashed lines show the theoretical sufficiency bounds.

configuration always admits a perturbation that results in the transformation of an object rank from k_x to k_t. Nevertheless, the theorem does illustrate an important trend: as the intrinsic dimensionality LID($\mathbf{x}$) increases, the minimum sufficiency bound on the perturbation proportion δ does tend to zero. In this section, we present several experiments in which the perturbation sufficiency bounds of Theorem 2 is contrasted with the empirical effect of perturbation on object rank.

Given a data set $S \subset \mathcal{S}$ of size n, an embedding dimension m, and a set of n_q query points, for each of the four cases described by the theorems, we record the minimum perturbation proportion δ applied to each query in order to achieve the following effects.

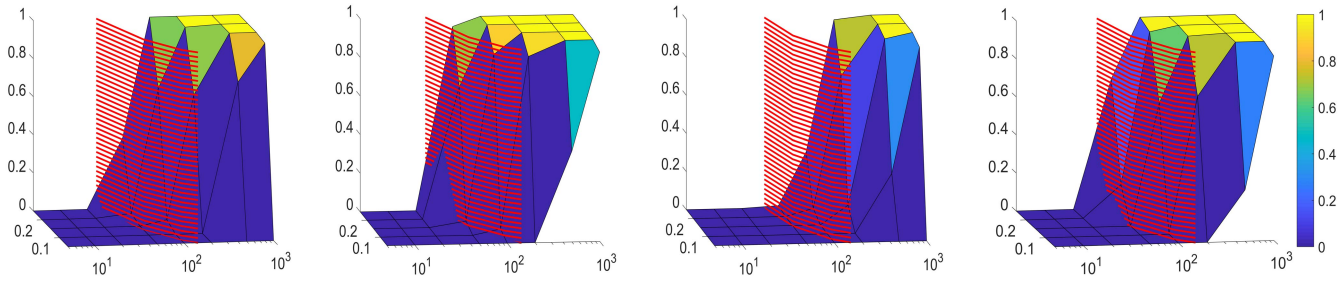
- *closer + query*: Move query point $\mathbf{x}$ *closer* to its k-NN $\mathbf{z}$, reducing the rank of $\mathbf{z}$ in the NN list from k to 1. This is an application of Theorem 2 where $\mathbf{x}$ is moved to query point $\mathbf{y}$ and the distributional rank is given by $F_{\mathbf{y}}$ (the *query-centric* setting).
- *closer + target*: Dataset point $\mathbf{x}$ being the k-NN of a targeted query $\mathbf{z}$, move $\mathbf{x}$ *closer* to the target reducing its rank in the NN list of $\mathbf{z}$ from k to 1. This is an application of Theorem 2 where the dataset point $\mathbf{x}$ is moved to $\mathbf{y}$ and the distributional rank is given by $F_{\mathbf{z}}$ (the *target-centric* setting).
- *farther + query*: Move query point $\mathbf{x}$ *farther* from its 1-NN $\mathbf{z}$, increasing the rank of $\mathbf{z}$ in the NN list from 1 to k (or above). This is an application where $\mathbf{x}$ is moved to query point $\mathbf{y}$ and the distributional rank is given by $F_{\mathbf{y}}$ (the *query-centric* setting).
- *farther + target*: Dataset point $\mathbf{x}$ being the 1-NN of a targeted query $\mathbf{z}$, move $\mathbf{x}$ *farther* from the target increasing the rank of $\mathbf{x}$ in the NN list of $\mathbf{z}$ from 1 to k (or above). This is an application of Theorem 2 where the dataset point $\mathbf{x}$ is moved to $\mathbf{y}$ and the distributional rank is given by $F_{\mathbf{z}}$ (the *target-centric* setting).

Our experimental results show a clear association between the LID at the query and the amount of perturbation.

A. Synthetic Data

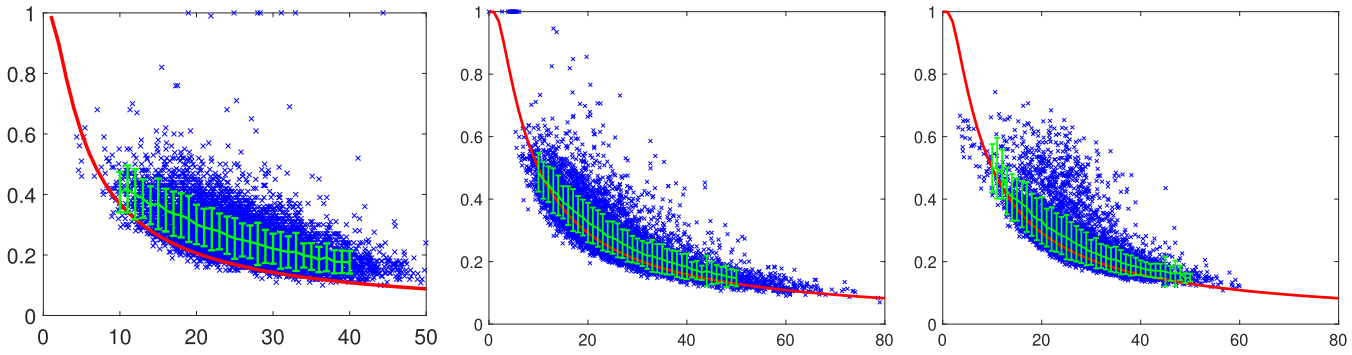
We consider a simple setting involving the standardized Gaussian (normal) distribution with i.i.d. components, from which we independently draw data sets with $n \in \{10^4, 10^5, \ldots, 10^9\}$ points, and varying dimensionality $m \in \{2, 5, 10, 20, 50, 100, 200, 500, 1000\}$. The normal distribution possesses the convenient property that the LID at each point is theoretically equal to the representational dimension m. Figures 3 and 4 show the empirically observed trends for $n_q = 100$ query points and $k = 1000$.

Figures 3 and 4 show the observed minimum δ averaged over all query points, for each of the four cases. Figure 3 plots this amount against the dimensionality m for each choice of n, while Figure 4 provides an alternative view of the same results by plotting the average minimum δ against n, for selected values of m. Two clear trends are noticeable: the observed minimum δ (i) decreases with LID($\mathbf{x}$), and (ii) decreases with n. For comparison, the theoretical bounds from Theorem 2 are also plotted (in red color). In Figure 3, the observed perturbation amounts are mostly above the theoretical sufficiency bounds, providing empirical support for the theorems, despite their being guaranteed to hold only asymptotically, as the data size n increases. On the other hand, the second trend is more clearly observed in Figure 4: for sufficiently large data size n, the empirical perturbation amounts do tend toward the theoretical sufficiency bounds (depicted with dashed lines).



(a) *Closer+query* w.r.t. $F_{\mathbf{y}_n}$. (b) *Closer+target* w.r.t. $F_{\mathbf{z}_n}$. (c) *Farther+query* w.r.t. $F_{\mathbf{y}_n}$. (d) *Farther+target* w.r.t. $F_{\mathbf{z}_n}$.

Fig. 5. Probability of attack success (color-coded) as a function of perturbation level δ (y-axis) and LID (x-axis) for $n = 10^6$ and $k = 1000$. The red surface shows the theoretical sufficiency bound.



(a) BIGANN SIFT1B $n = 1.0 \times 10^9$, $k = 100$ (b) ImageNet $n = 1.3 \times 10^6$, $k = 1,000$ (c) CIFAR-10 $n = 5.0 \times 10^4$, $k = 1,000$

Fig. 6. Experiments on three real data sets, plotting the minimum perturbation required δ (y-axis) vs. estimated LID (x-axis). Red curve: theoretical bound from Theorem 2. Green bars: empirical mean and standard deviation.

The last experiment focuses on an adversarial retrieval scenario. The adversary applies a fixed perturbation level δ and the experiment records whether the goal of the attack is achieved. For the *closer* scenario, the attack succeeds when the k-NN becomes the closest neighbor. For the *farther* scenario, the attack succeeds when the rank of the first neighbor becomes greater or equal to k. Figure 5 shows a surface plot of the probability of success (average over n_q queries) as a function of δ and d for $n = 10^6$ and $k = 1000$. The plot shows a phase transition in which the probability of success quickly goes from 0 to 1 at a critical relationship between δ and LID, at values that exceed the sufficiency bound (the surface shown in red). Again, we observe the same trend: the critical degree of perturbation decreases as LID increases.

B. Real Data

On real data we illustrate the *closer* + *target* case through experiments that (i) support the asymptotic behavior of Theorem 2 with distributional rank $F_{\mathbf{y}}$ when n is extremely large, and (ii) demonstrate that δ decreases as LID increases. LID values were obtained using the maximum likelihood estimator described by [36], computed from the distances of the 100 nearest neighbors for each query point.

Figure 6(a) plots the values for perturbation δ when using the BIGANN_SIFT1B data set [41], where $d = 128$ and $n = 10^9$. Here, we chose $n_q = 10,000$ and $k = 100$. In order to estimate the mean minimum perturbation, we group the LID values into integer bins. The mean and standard deviation of the perturbation levels for each bin is reported in green in this figure, whenever there are sufficiently many samples to compute reliable statistics. In this experiment, n is extremely large, revealing the asymptotic behavior of Theorem 2. Very few values for δ are below the theoretical sufficiency bound (only 2.2% of the query points). This experiment uses descriptors with estimated LID in the low to moderate range, smaller than the space dimension m. In contrast, Figures 6(b) and 6(c) show complementary configurations where estimated LID are larger, but still much smaller than the space dimension m. Another difference is that n is much smaller for these datasets. Figure 6(b) plots δ against LID for the ImageNet data set [42]. This data set 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 and $k = 1{,}000$. Figure 6(c) corresponds to the case of the CIFAR-10 data set [43], which consists of $n = 50{,}000$ training images. 10,000 test images are also provided, which we use as queries with $k = 1{,}000$. Both data sets are fed into a deep neural network to extract high level features. Specifically, we extract $m = 2048$ features from the global average pooling layer in the Resnet-50 network [44].

As can be seen from Figures 6(b) and 6(c), the observed amount of perturbation decreases as the LID grows. As expected, the theoretical curves pass through the data clouds because n is too small for the asymptotic trends to fully assert themselves: around 25% of the query points have an empirical minimum perturbation smaller than the asymptotical lower bound.

V. Conclusion

This paper has presented a theoretical explanation of the effect of adversarial perturbation on nearest neighborhoods under the Euclidean distance metric: the larger the local intrinsic dimensionality and data-set size, the smaller the amount of adversarial perturbation required to transition between 1-NNs and k-NNs (by expectation). These theoretical trends are confirmed experimentally for synthetic and real data sets. Our results demonstrate 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, the direction of perturbation (*closer* or *farther*) and vantage point (*query-centric* or *target-centric*).

The traditional view of dimensionality in practice has been to treat it as a parameter of the entire representation space. Our analysis in terms of LID shows that the effects of dimensionality are not necessarily dependent of the properties of the representation space. In situations where the dimensional characteristics are observed to vary from locality to locality, our analysis underscores the need for applications of perturbation — including the defense and detection of adversarial attack — to take the variability of local intrinsic dimensionality into account.

A. Adversarial Retrieval

The impact of this article on adversarial retrieval is straightforward where vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ correspond to feature vectors of pieces of content. However, this theoretical paper does not consider the final step of a practical attack, which is the inversion of the feature extraction: the crafting of a perturbed content object whose feature vector maps to the vector $\mathbf{y}$ as defined in Section III. The generation of perturbed objects is equivalent to a local inversion of the feature extraction operation. From original content producing a feature vector $\mathbf{x}$, we look for perceptually similar content whose feature vector is $\mathbf{y}$, provably very close to $\mathbf{x}$. Previous research has addressed this issue in the context of image retrieval based on local descriptors [21]–[23].

Current state-of-the-art image search frameworks work with global descriptors extracted by deep neural networks. The local inversion of this feature extraction is feasible in practice due to the existence of efficient back-propagation mechanisms [2], [45]. Very recent works apply such local inversion techniques in the context of adversarial image retrieval, where perturbed query images have been generated that are visually or semantically close to their corresponding original images [46], [47].

B. Adversarial Classification

Our analysis also has implications for adversarial classification — in particular, we have proven strong theoretical statements concerning the effect of perturbation on 1-NN classification. The 1-NN classifier has long been known to be 'asymptotically optimal,' in that the probability of error is bounded from above by twice the Bayes error, as the training set size tends to infinity [48], [49]. In this sense, half the classification information in an infinite sample set can be regarded as residing with the nearest neighbor of the test item.

Within Euclidean space or other vector spaces, 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 into 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 perturb 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.

The question remains open as to whether a quantitative explanation analogous to those of our theorem 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 1 is likely to hold, provided that a transformation exists between the original domain and an appropriate Euclidean domain. The theorems 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. This is consistent with recently-proposed uses of LID for the characterization of adversarial examples for deep learning classification [50], where it was found that adversarial examples tend to be associated with higher LID estimates. Reference [51] has questioned whether high LID is a property of all adversarial examples, or whether it is a side effect of particular types of attacks. 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.

Acknowledgment

The majority of his work on this article was done while Xuan Vinh Nguyen was with the University of Melbourne, Australia.

References

[1] D. Lowd and C. Meek, "Adversarial learning," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2005, pp. 641–647.

[2] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Apr. 2014.

[3] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 1625–1634. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Eykholt_Robust_Physical-World_Attacks_CVPR_2018_paper.html, doi: 10.1109/CVPR.2018.00175.

[4] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 1528–1540.

[5] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognit.*, vol. 84, pp. 317–331, 2018, doi: 10.1016/j.patcog.2018.07.023.

[6] A. Fawzi, H. Fawzi, and O. Fawzi, "Adversarial vulnerability for any classifier," in *Proc. 31st Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montréal, QC, Canada, Dec. 2018, pp. 1186–1195. [Online]. Available: http://papers.nips.cc/paper/7394-adversarialvulnerability-for-any-classifier

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[8] N. Papernot, P. D. McDaniel, and I. J. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," *CoRR*, vol. abs/1605.07277, 2016. [Online]. Available: http://arxiv.org/abs/1605.07277

[9] F. Tramèr, N. Papernot, I. J. Goodfellow, D. Boneh, and P. D. McDaniel, "The space of transferable adversarial examples," *CoRR*, vol. abs/1704.03453, 2017. [Online]. Available: http://arxiv.org/abs/1704.03453

[10] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially robust generalization requires more data," in *Proc. 31st Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montréal, QC, Canada, Dec. 2018, pp. 5019–5031. [Online]. Available: http://papers.nips.cc/paper/7749-adversariallyrobust-generalization-requires-more-data

[11] A. Fawzi, O. Fawzi, and P. Frossard, "Analysis of classifiers' robustness to adversarial perturbations," *Mach. Learn.*, vol. 107, no. 3, pp. 481–508, 2018, doi: 10.1007/s10994-017-5663-3.

[12] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet, "Adversarial manipulation of deep representations," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, San Juan, PR, USA, May 2016. [Online]. Available: http://arxiv.org/abs/1511.05122

[13] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard, "Robustness of classifiers: From adversarial to random noise," in *Proc. 29th Annu. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 1624–1632. [Online]. Available: http://papers.nips.cc/paper/6331-robustness-ofclassifiers-from-adversarial-to-random-noise

[14] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, and S. Soatto, "Analysis of universal adversarial perturbations," *CoRR*, vol. abs/1705.09554, 2017. [Online]. Available: http://arxiv.org/abs/1705.09554

[15] E. D. Cubuk, B. Zoph, S. S. Schoenholz, and Q. V. Le, "Intriguing properties of adversarial examples," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr./May 2018. [Online]. Available: https://openreview.net/forum?id=Skz1zaRLz

[16] T. Tanay and L. D. Griffin, "A boundary tilting persepective on the phenomenon of adversarial examples," *CoRR*, vol. abs/1608.07690, 2016. [Online]. Available: http://arxiv.org/abs/1608.07690

[17] J. Gilmer *et al.*, "Adversarial spheres," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr./May 2018. [Online]. Available: https://openreview.net/forum?id=SkthlLkPf

[18] Y. Wang, S. Jha, and K. Chaudhuri, "Analyzing the robustness of nearest neighbors to adversarial examples," in *Proc. 35th Int. Conf. Mach. Learn. (PMLR)*, J. Dy and A. Krause, Eds., vol. 80, 2018, pp. 5133–5142.

[19] N. Papernot and P. D. McDaniel, "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning," *CoRR*, vol. abs/1803.04765, 2018. [Online]. Available: http://arxiv.org/abs/1803.04765

[20] C. Sitawarin and D. A. Wagner, "On the robustness of deep K-nearest neighbors," in *Proc. IEEE Secur. Privacy Workshops (SP)*, San Francisco, CA, USA, May 2019, pp. 1–7, doi: 10.1109/SPW.2019.00014.

[21] T.-T. Do, E. Kijak, T. Furon, and L. Amsaleg, "Challenging the security of content-based image retrieval systems," in *Proc. IEEE Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2010, pp. 52–57.

[22] T. Do, E. Kijak, L. Amsaleg, and T. Furon, "Security-oriented picture-in-picture visual modifications," in *Proc. Int. Conf. Multimedia Retr. (ICMR)*, 2012, p. 13.

[23] T.-T. Do, E. Kijak, L. Amsaleg, and T. Furon, "Enlarging hacker's toolbox: Deluding image recognition by attacking keypoint orientations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 1817–1820.

[24] R. Feng and B. Prabhakaran, "Facilitating fashion camouflage art," in *Proc. 21st ACM Int. Conf. Multimedia (MM)*, 2013, pp. 793–802.

[25] C. Chen, A. Dantcheva, T. Swearingen, and A. Ross, "Spoofing faces using makeup: An investigative study," in *Proc. IEEE Int. Conf. Identity, Secur. Behav. Anal. (ISBA)*, Feb. 2017, pp. 1–8.

[26] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang, "Invisible mask: Practical attacks on face recognition with infrared," *CoRR*, vol. abs/1803.04683, 2018. [Online]. Available: http://arxiv.org/abs/1803.04683

[27] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *J. Mach. Learn. Res.*, vol. 11, pp. 2487–2531, 2010.

[28] L. Amsaleg *et al.*, "The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality," in *Proc. IEEE Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2017, pp. 1–6.

[29] M. E. Houle, "Dimensionality, discriminability, density & distance distributions," in *Proc. 13th IEEE Int. Conf. Data Mining Workshops (ICDMW)*, 2013, pp. 468–473.

[30] M. E. Houle, "Local intrinsic dimensionality I: An extreme-value-theoretic foundation for similarity applications," in *Proc. Int. Conf. Similarity Search Appl. (SISAP)*, 2017, pp. 64–79.

[31] M. E. Houle, "Local intrinsic dimensionality II: Multivariate analysis and distributional support," in *Proc. Int. Conf. Similarity Search Appl. (SISAP)*, 2017, pp. 80–95.

[32] A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein, "Are adversarial examples inevitable?" in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, May 2019. [Online]. Available: https://openreview.net/forum?id=r1lWUoA9FQ

[33] S. Mahloujifar, D. I. Diochnos, and M. Mahmoody, "The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI), 31st Innov. Appl. Artif. Intell. Conf. (IAAI), 9th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, Honolulu, HI, USA, Jan./Feb. 2019, pp. 4536–4543, doi: 10.1609/aaai.v33i01.33014536.

[34] M. E. Houle, H. Kashima, and M. Nett, "Generalized expansion dimension," in *Proc. IEEE 12th Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2012, pp. 587–594.

[35] D. R. Karger and M. Ruhl, "Finding nearest neighbors in growth-restricted metrics," in *Proc. 34th Annu. ACM Symp. Theory Comput. (STOC)*, 2002, pp. 741–750.

[36] L. Amsaleg *et al.*, "Extreme-value-theoretic estimation of local intrinsic dimensionality," *Data Mining Knowl. Discovery*, vol. 32, no. 6, pp. 1768–1805, Nov. 2018.

[37] R. Huisman, K. G. Koedijk, C. J. M. Kool, and F. Palm, "Tail-index estimates in small samples," *J. Bus. Econ. Statist.*, vol. 19, no. 2, pp. 208–216, Apr. 2001.

[38] M. I. Gomes, L. Canto E Castro, M. I. F. 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, no. 1, pp. 3–34, Mar. 2008.

[39] L. Amsaleg, O. Chelly, M. E. Houle, K. Kawarabayashi, M. Radovanović, and W. Treeratanajaru, "Intrinsic dimensionality estimation within tight localities," in *Proc. 19th SIAM Int. Conf. Data Mining (SDM)*, 2019, pp. 37–65.

[40] O. B. Allen, "Quantiles," in *Encyclopedia of Environmetrics*. Hoboken, NJ, USA: Wiley, 2006.

[41] H. Jégou, R. Tavenard, M. Douze, and L. Amsaleg, "Searching in one billion vectors: Re-rank with source coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 861–864.

[42] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[43] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Univ. Toronto, Toronto, ON, Canada, 2009.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[45] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 427–436.

[46] Z. Liu, Z. Zhao, and M. A. Larson, "Who's afraid of adversarial queries?: The impact of image modifications on content-based image retrieval," in *Proc. Int. Conf. Multimedia Retr. (ICMR)*, Ottawa, ON, Canada, Jun. 2019.

[47] G. Tolias, F. Radenovic, and O. Chum, "Targeted mismatch adversarial attack: Query with a flower to retrieve the tower," in *Proc. ICCV*, 2019, pp. 5037–5046.

[48] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
[49] C. J. Stone, "Consistent nonparametric regression," *Ann. Statist.*, vol. 5, no. 4, pp. 595–645, 1977.
[50] X. Ma *et al.*, "Characterizing adversarial subspaces using local intrinsic dimensionality," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–15.
[51] A. Athalye, N. Carlini, and D. A. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, Jul. 2018, pp. 274–283.

Laurent Amsaleg (Member, IEEE) received the Ph.D. degree from the University of Paris 6. He is currently a Senior Researcher with CNRS. He also leads the Linkmedia Research Group, IRISA/INRIA Laboratory, Rennes, France. His research interests include high-dimensional indexing at scale and multimedia analytics, as well as the many facets of the security issues in relation with the processing of extremely large collections of multimedia material. Topics dealing with privacy and adversarial machine learning are therefore central to his work.

James Bailey (Member, IEEE) was an Australian Research Council Future Fellow. He is currently a Professor with the Melbourne School of Engineering, The University of Melbourne, and the Program Lead of Artificial Intelligence. He is also a Researcher in the field of machine learning and artificial intelligence, including interdisciplinary applications and operational frameworks. His interests particularly relate to the assurance, certification, and safety of systems based on machine learning and artificial intelligence.

Amélie Barbe received the master's degree in computer science from ENS de Rennes in 2018. She is currently pursuing the Ph.D. degree in optimal transport and distances between attributed graphs with ENS de Lyon.

Sarah M. Erfani is currently a Senior Lecturer with the School of Computing and Information Systems, The University of Melbourne. Her research interests include machine learning, large-scale data mining, cyber security, and data privacy.

Teddy Furon (Senior Member, IEEE) received the M.Sc. degree and the Ph.D. degree in signal processing from Telecom ParisTech, in 1998 and 2002, respectively. His field of interest is the security related to multimedia and signal processing. He has worked both in industry (Thomson, Technicolor) and academia (Université Catholique de Louvain, Belgium, and now Inria Rennes, France, in the Linkmedia team). He co-founded of the company LAMARK protecting rights of photo agencies. He has been an Associate Editor of four journals, including the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.

Michael E. Houle (Member, IEEE) received the Ph.D. degree in computational geometry from McGill University in 1989. He has held positions in Japan at Kyushu University and the University of Tokyo, and in Australia at the University of Newcastle and the University of Sydney. From 2001 to 2004, he was a Visiting Scientist with the IBM Japan's Tokyo Research Laboratory. His research interests include algorithmics, data structures, relational visualization, machine learning, and data mining, particularly as regards issues involving dimensionality and scalability in the context of search, clustering, classification, and outlier detection. He is currently a Visiting Professor with the National Institute of Informatics (NII), Japan.

Miloš Radovanović is currently an Associate Professor with the Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad. His research interests span many areas of machine learning and data mining, with special focus on problems related to high-data dimensionality, complex networks, time-series analysis, text mining, techniques for classification, clustering, and outlier detection. He is a Managing Editor of the *Computer Science and Information Systems Journal* (ComSIS).

Xuan Vinh Nguyen received the Ph.D. degree from the University of New South Wales, Australia. He held research positions at Monash and Melbourne universities. He is currently a Deep Learning Engineer and a Data Scientist with NVIDIA, having published more than 50 scientific articles attracting more than 2500 citations. This work was done a large part while he was with the University of Melbourne. At NVIDIA, his work spans a wide range of deep learning and AI applications, including speech, language and vision processing, and recommender systems.