Intrinsic Geometric Vulnerability of High-Dimensional Artificial Intelligence

Luca Bortolussi[a] **and Guido Sanguinetti**[b]

[a]Department of Mathematics and Geosciences, University of Trieste; [b]School of Informatics, University of Edinburgh

This manuscript was compiled on November 9, 2018

The success of modern Artificial Intelligence (AI) technologies depends critically on the ability to learn non-linear functional dependencies from large, high dimensional data sets. Despite recent high-profile successes, empirical evidence indicates that the high predictive performance is often paired with low robustness, making AI systems potentially vulnerable to adversarial attacks. In this report, we provide a simple intuitive argument suggesting that high performance and vulnerability are intrinsically coupled, and largely dependent on the geometry of typical, high-dimensional data sets. Our work highlights a major potential pitfall of modern AI systems, and suggests practical research directions to ameliorate the problem.

Artificial Intelligence | Adversarial Attacks | High-dimensional geometry | Computer Security

Artificial Intelligence (AI) is colonising all areas of human endeavour, and its impact is widely predicted to grow exponentially in the next decades. Techniques such as deep learning have significantly improved on the state of the art in areas as diverse as computer vision, speech recognition and medical imaging (1–5), and have already reached super-human performance in games such as GO and classical ATARI video games (6, 7). Buoyed by these successes, many researchers are heralding a new golden age for AI, and many governments and major corporations have started multi-billion dollar research investments in the development and application of AI.

Despite these undeniable achievements, the mathematical and statistical bases for AI's success, and consequently its general applicability, remain largely unclear. Techniques such as deep learning work by defining a broad class of possible input/output functions underpinning the structure of the data. Such functional classes are encoded in the network structure, and in the so called *activation functions*, and are usually sufficiently general as to approximate arbitrarily well any smooth function. The specific predictive function is chosen by optimising a measure of fit to a subset of the data (*training data*), and performance is evaluated statistically over a held out subset of the data (*test set*). The training procedure (learning) is normally some variation of (stochastic) gradient descent, and much of deep learning research is concerned with the development of heuristic methods to improve the learning procedure or with the engineering of network architectures tailored to specialized tasks. The success of this approach has largely taken by surprise even the practitioners: deep learning methods were essentially already well known in the eighties, and were largely abandoned in the intervening time as too complex and prone to overfitting.

Some attribute the new found success of deep learning methods to a combination of more powerful hardware and, crucially, much larger data sets that have become available following the advent of the internet and social networks. Recent studies on simplified models have shown how the optimisation problem itself (a notorious stumbling block for early generations of deep learning) may become simpler in the large-data regime (8–13). However, this explanation is still unsatisfactory: it is well known that approximating a Lipschitz continuous function to a fixed precision requires a number of instances that grows exponentially with the dimension of the function's domain (e.g. (14)). AI methods routinely provide excellent performance on very high-dimensional ($\sim 10^4$) data sets consisting of a few million examples. These numbers may seem very large, but, in terms of learning general functions of tens of thousands of variables, they are not.*

A second, less widely known limitation of deep AI methodologies is their vulnerability to adversarial attacks. As early as 2013 (16), researchers observed that minimal perturbations to test data could completely overturn the prediction of a deep

*Images from the ImageNet dataset (15), typically used to train deep Convolutional Neural Network for classification, have a working resolution of 256x256 pixels, with three channels, which amounts to an input space of about $n = 195,000$ dimensions. Approximating a Lipschitz function with Lipschitz constant L with error ϵ requires $O((\frac{L}{\epsilon})^n)$ points, which is a super-astronomical number even for relatively large ϵ.

Significance Statement

Modern artificial intelligence (AI) is critically reliant on learning functions from high dimensional inputs such as images and speech. Recent research has shown that, while powerful in terms of prediction, such techniques are often vulnerable to adversarial attacks. In this paper, we argue that such vulnerability is intrinsic for a wide class of AI systems. These results have important implications about our ability to build secure AI systems.

Both authors conceived and carried out the research, and wrote the paper.

No conflicts of interest.

[2]To whom correspondence should be addressed. E-mail: gsanguin@inf.ed.ac.uk

learning algorithm. For example, in a computer vision application, flipping a suitably chosen single pixel in a (correctly classified) image of a dog could return a prediction of a cat (17). While this observation did not stop the onward march of AI (even in safety critical applications such as self-driving cars), no effective solutions to the problem of adversarial vulnerability of deep learning methods have been found. A competition held at the last edition of the premier machine learning conference NIPS provided some promising preliminary results (18), but unfortunately further work (19) later showed that even these defences could be broken with a stronger attack strategy.

In this brief report, we take an alternative, geometric perspective to analyse the performance of AI methods on high-dimensional data sets. We focus on the simple case of binary classification: the prediction task consists of assigning a binary label to points in a high dimensional vector space (which we will take to be $\mathbb{R}^N$ for simplicity), based on a training set of labelled instances. Extension to multi-class classification problems is trivial. Our arguments show that indeed complex high dimensional classifiers can perform well only when the data distribution exhibit some special properties. Additionally, we show that vulnerability is a *direct consequence* of the structures that make learning successful, and therefore the inevitable other side of the performance medal. We focus on providing intuitive arguments that can cover the general situations, rather than rigour; proofs would be difficult to provide without strong simplifying assumptions, and would not necessarily add to our understanding of the root causes of the problem.

Results

To make progress, we start by introducing the concept of a *locally complex classifier*. Let $\mathcal{D}$ be a data set consisting of input/ output pairs $\{\mathbf{x}_i, y_i\}$, assumed to be drawn i.i.d. from an (unknown) distribution $p(\mathbf{x}, y)$. Input variables $\mathbf{x}$ are points in a high-dimensional vector space $\mathbf{x} \in \mathbb{R}^N$, with N very large, while outputs y are binary labels. A classifier is therefore a map $C\colon \mathbb{R}^{\mathbb{N}} \to \{-1, 1\}$ assigning to each point in input space a binary label. We will assume all classifiers to be locally constant functions, meaning that, for almost every point in input space classified as 1 (resp. -1), there exists a finite neighbourhood where the classifier does not change value. The *discriminant* d_C defined by the classifier C is the boundary in $\mathbb{R}^N$ of the pre-image of the value 1 (or equivalently -1); from the local constancy assumption, it follows that the discriminant is a set of measure 0, and defines a surface within $\mathbb{R}^N$. The discriminant surface is the central object of study in this paper; the following definition allows us to reason precisely on the complexity of the discriminant.

Definition. A classifier $C\colon \mathbb{R}^{\mathbb{N}} \to \{-1, 1\}$ is *locally complex* at $\mathbf{x}^*$ if the discriminant d_C near $\mathbf{x}^*$ can be well approximated locally by a set of $\Omega(N)$ independent linear equations $\mathbf{w}_i^T \mathbf{x} + c = 0$.

Intuitively, this definition captures the complexity of the discriminant by trying to quantify its "wriggliness" in high dimensions (see Figure 1), requiring the discriminant to be defined by a number of linear equalities of order N. Locally complex classifiers include fully grown decision trees, and deep neural networks with large numbers of nodes; in particular, deep networks using the popular rectified linear units (ReLU) activation function partition the input space in a large number of polyhedra (exponential in the number of layers) so that they are locally complex at very many points. Linear classifiers, on the other hand, express their discriminant as a single inequality, and are therefore not complex anywhere (as is to be expected).

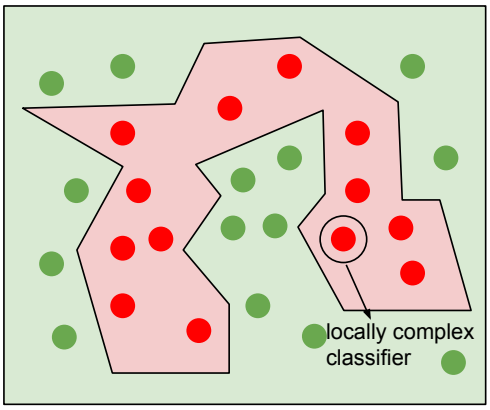

Fig. 1. Schematic example of locally complex discriminant

What can the local geometry of the discriminant tell us on the performance of the classifier? Complex predictors in machine learning are often associated with overfitting problems, and indeed the following observation suggests that this problem, under certain conditions, affects all locally complex classifiers.

Observation. Let $\mathbf{x}_i$ be a training point where the classifier is locally complex, and let the (class conditional) data generating distribution be non-degenerate and with unbounded support in all directions. Then, with high probability, nearby points drawn from the data generating distribution will be misclassified.

This follows simply from the fact that generating a nearby point is equivalent to sampling a "noisy version" of the training point, and since the noise is unbounded in all directions the probability that in at least one of the $\Omega(N)$ "fragile" directions (i.e. those defining the discriminant) we sample a value that crosses the boundary grows to one exponentially in N. We remark that unbounded support is a very common assumption for a noise model; for example, the multivariate Gaussian distribution has support over the whole of $\mathbb{R}^N$.

This observation implies that, if a locally complex classifier performs empirically well in high dimensions, then the true data distribution must concentrate. In other words, test points

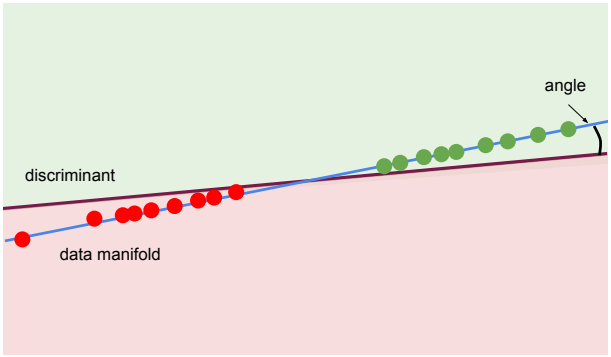

Fig. 2. Schematic exemplification of fragility of linear classifiers in high-dimension.

must lay on a subset of the input space of very small dimension. This observation chimes with many intuitive explanations proffered in recent years for the success of deep learning, which variously remarked on the high degree of symmetry of natural images, or on the equivalence of many local optima of the networks. In the following, we will assume that the support of the data distribution lies exactly on a low-dimensional submanifold of the ambient input space, of dimension $M \ll N$.

The true low-dimensionality of the data directly solves the conundrum of function approximation in high dimension: approximating a function requires data sets of exponentially increasing size only if the function is genuinely defined on a high dimensional domain. If all we need is a good approximation on a very small subset of the space, then the problem no longer arises. Still, the result of learning a high-dimensional classifier is a function defined on the whole ambient space.

What will this function look like outside of the constrained data manifold? The precise answer will depend on many factors, including the training procedure and data, yet we can safely assume that it will be essentially random once sufficiently far from the data manifold. And if the data manifold is genuinely concentrated in low dimensions and embedded in a high dimensional space, sufficiently far might actually mean very near. To see why, consider the following simple example.

Example: a linear classifier for apparently high dimensional data. Consider a data generating distribution whose class-conditionals are well-separated Gaussians in $M \ll N$ dimensions (see Figure 2 for N=2, M=1). Let us use logistic regression (LR) to classify this data; LR defines a hyperplane as a discriminant, and therefore requires the specification of a bias vector (N parameters) and an orthogonal unit vector ($N-1$ parameters). Since the data is well separated, LR will find very accurately the optimal $M-1$ dimensional hyperplane in the data space, constraining $N+M-1$ parameters. The remaining $N-M$ parameters are unconstrained, and their value will be essentially random. If we interpret the unconstrained parameters as azimuth angles, then the distance from any data point to the discriminant will be proportional to the sine of one such angles. If N is very large and $N-M$ is $O(N)$, the probability that at least one angle, hence the distance of a training point from the discriminant, will be smaller than a constant ϵ will approach 1 exponentially, therefore showing that this classifier is fragile by construction.

This property is essentially equivalent to other observations in literature about lack of robustness of linear classifiers, though it has a clearer geometric flavour. For instance, in (20), the authors observe that principal components corresponding to small eigenvalues can have an associated high weight of the linear classifier. This can be seen as the counterpart of having low angular coordinates. This clearly explains why linear attacks are easy to find in high-dimensional models (21), particularly when the data manifold has a much lower dimensionality: a step away from the manifold will typically involve a linear combination of several directions normal to the manifold having large weights, resulting in a big change in the linear classifier. Moreover, several such directions will retain large weights also when learned on a different dataset, as the weights are assigned randomly, showing that linear attacks are likely to generalise (21).

Notice that Logistic Regression is not a locally complex classifier. Indeed, under some simplifying conditions, (22) recently proved that *any* classifier in high dimensions is vulnerable when the data distribution is low dimensional. Geometry in high-dimensional spaces has also been recently advocated as a possible cause for adversarial attacks in (23), where authors study a highly idealised scenario in which two-class data is distributed in two concentric spheres, and observe that misclassified points tend to appear on average close to any test point, with a distance decreasing with the square root of the dimension. A similar result, in a more general setting of a two classes problem on a sphere or a unit cube, has been discussed very recently in (24), leveraging specialised isoperimetric inequalities and connecting it to some geometric properties connected with the the data manifold.

An additional intriguing feature is that many adversarial examples (i.e. small perturbations almost indistinguishable by humans that fool deep classifiers) generalise to different architectures, possibly trained on different datasets (21, 25). Also in this case, the geometry of the data manifold and its embedding in a high dimensional space are likely to be involved. On the one hand, one can find directions that generate examples which are sufficiently far from the data manifold, the so called linear attacks in (21). The high dimensionality of the input space *de facto* implies that such directions exist and are common, also for simple linear models. However, not all adversarial examples are of this category, and the low dimensional data manifold itself is likely to be intrinsically complex once embedded into a high-dimensional space. This intuitively means that each point in the data manifold is likely to be close to other parts of the manifold corresponding to different classes. Evidence in this direction comes from the fact that adversarial examples have been found to typically have a higher local

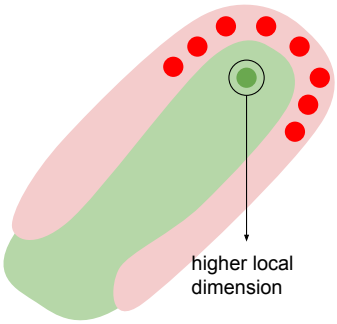

Fig. 3. Schematic example of increase in intrinsic local dimension.

intrinsic dimensionality than training points (26, 27), suggesting that robust adversarial examples are found in directions in space where the data manifold folds and has a more complex local geometry (see Figure 3).

Discussion.

In summary, our results recapitulate a number of previous observations that were broadly conjectured in the technical community, bringing them together under a novel, intuitive geometric perspective. A major new insight arising from this perspective is that complex classifiers can only work well in circumstances where they necessarily are vulnerable.

Our work also illustrates some possible directions to ameliorate the problem. Several groups are already investigating the possibility of adding local consistency constraints to the objective function of a neural network classifier (21, 25, 28–31) for example in the form of ∞-norm robustness. Such approaches show promise, yet, in order not to compromise performance, only very light regularisation can be applied. An alternative is to avoid the high-dimensionality trap by pre-processing data with a dimensionality reduction technique (32, 33). Such an approach is appealing, as it may greatly simplify the data manifold geometry, and in some simple cases can be analysed theoretically (34), though it may still be vulnerable to white box adversarial attacks when combined in a pipeline with a (complex) classifier (24). Notions of intrinsic dimensionality (27, 35) may play a useful role in understanding vulnerability, and indeed PCA has been advocated as a defense strategy against adversarial attacks (20). A different direction would be to adopt a Bayesian perspective, provided we can tackle its formidable computational challenges: this would both regularise unconstrained directions and, by quantifying posterior uncertainty on model parameters, would potentially automatically detect vulnerable directions.

1. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *arXiv:1512.00567 [cs]*, December 2015. URL http://arxiv.org/abs/1512.00567. arXiv: 1512.00567.
2. Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, December 2017. ISSN 1361-8415. . URL http://www.sciencedirect.com/science/article/pii/S1361841517301135.
3. Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, September 2014. URL http://arxiv.org/abs/1409.1556. arXiv: 1409.1556.
4. Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent Trends in Deep Learning Based Natural Language Processing. *arXiv:1708.02709 [cs]*, August 2017. URL http://arxiv.org/abs/1708.02709. arXiv: 1708.02709.
5. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012. ISSN 1053-5888. .
6. Volodymyr Mnih Koray Kavukcuoglu, David Silver Alex Graves, Ioannis Antonoglou, and Daan Wierstra Martin Riedmiller. Playing Atari with Deep Reinforcement Learning. page 9, 2013.
7. David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, October 2017. ISSN 0028-0836, 1476-4687. . URL http://www.nature.com/doifinder/10.1038/nature24270.
8. Grzegorz Swirszcz, Wojciech Marian Czarnecki, and Razvan Pascanu. Local minima in training of neural networks. *arXiv:1611.06310 [cs, stat]*, November 2016. URL http://arxiv.org/abs/1611.06310. arXiv: 1611.06310.
9. Levent Sagun, V. Ugur Guney, Gerard Ben Arous, and Yann LeCun. Explorations on high dimensional landscapes. *arXiv:1412.6615 [cs, stat]*, December 2014. URL http://arxiv.org/abs/1412.6615. arXiv: 1412.6615.
10. Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The Loss Surfaces of Multilayer Networks. *arXiv:1412.0233 [cs]*, November 2014. URL http://arxiv.org/abs/1412.0233. arXiv: 1412.0233.
11. Ian J. Goodfellow, Oriol Vinyals, and Andrew M. Saxe. Qualitatively characterizing neural network optimization problems. *arXiv:1412.6544 [cs, stat]*, December 2014. URL http://arxiv.org/abs/1412.6544. arXiv: 1412.6544.
12. C. Daniel Freeman and Joan Bruna. Topology and Geometry of Half-Rectified Network Optimization. *arXiv:1611.01540 [cs, stat]*, November 2016. URL http://arxiv.org/abs/1611.01540. arXiv: 1611.01540.
13. Carlo Baldassi and Riccardo Zecchina. Efficiency of quantum vs. classical annealing in nonconvex learning problems. *Proceedings of the National Academy of Sciences*, page 201711456, 2018.
14. Joseph F Traub. *Information-based complexity*. John Wiley and Sons Ltd., 2003.
15. ImageNet, http://www.image-net.org/. URL http://www.image-net.org/.
16. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
17. Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi. One pixel attack for fooling deep neural networks. *arXiv:1710.08864 [cs, stat]*, October 2017. URL http://arxiv.org/abs/1710.08864. arXiv: 1710.08864.
18. NIPS 2017: Non-targeted Adversarial Attack, https://www.kaggle.com/c/nips-2017-non-targeted-adversarial-attack, 2017. URL https://www.kaggle.com/c/nips-2017-non-targeted-adversarial-attack.
19. Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial Risk and the Dangers of Evaluating Against Weak Attacks. *arXiv:1802.05666 [cs, stat]*, February 2018. URL http://arxiv.org/abs/1802.05666. arXiv: 1802.05666.
20. Arjun Nitin Bhagoji, Daniel Cullina, Chawin Sitawarin, and Prateek Mittal. Enhancing Robustness of Machine Learning Systems via Data Transformations. *arXiv:1704.02654 [cs]*, April 2017. URL http://arxiv.org/abs/1704.02654. arXiv: 1704.02654.
21. Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv:1412.6572 [cs, stat]*, December 2014. URL http://arxiv.org/abs/1412.6572. arXiv: 1412.6572.
22. Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. *arXiv:1802.08686 [cs, stat]*, February 2018. URL http://arxiv.org/abs/1802.08686. arXiv: 1802.08686.
23. Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial Spheres. *arXiv:1801.02774 [cs]*, January 2018. URL http://arxiv.org/abs/1801.02774. arXiv: 1801.02774.
24. Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv:1809.02104 [cs, stat]*, September 2018. URL http://arxiv.org/abs/1809.02104.
25. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks.

arXiv:1312.6199 [cs], December 2013. URL http://arxiv.org/abs/1312.6199. arXiv: 1312.6199.

26. Laurent Amsaleg, James Bailey, Dominique Barbe, Sarah Erfani, Michael E. Houle, Vinh Nguyen, and Milos Radovanovic. The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality. pages 1–6. IEEE, December 2017. ISBN 978-1-5090-6769-5. . URL http://ieeexplore.ieee.org/document/8267651/.
27. Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey. Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality. *arXiv:1801.02613 [cs]*, January 2018. URL http://arxiv.org/abs/1801.02613. arXiv: 1801.02613.
28. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *NIPS*, page 9, 2014.
29. Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. *arXiv:1511.04508 [cs, stat]*, November 2015. URL http://arxiv.org/abs/1511.04508. arXiv: 1511.04508.
30. Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the Science of Security and Privacy in Machine Learning. *arXiv:1611.03814 [cs]*, November 2016. URL http://arxiv.org/abs/1611.03814. arXiv: 1611.03814.
31. Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. *arXiv:1608.04644 [cs]*, August 2016. URL http://arxiv.org/abs/1608.04644. arXiv: 1608.04644.
32. Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.
33. Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.
34. Timothy I Cannings and Richard J Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035, 2017.
35. Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, September 2017. ISSN 2045-2322. . URL https://www.nature.com/articles/s41598-017-11873-y.