The Intrinsic Dimension of Images and Its Impact on Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

It is widely believed that natural image data exhibits low-dimensional structure despite being embedded in a high-dimensional pixel space. This idea underlies a common intuition for the success of deep learning and has been exploited for enhanced regularization and adversarial robustness. In this work, we apply dimension estimation tools to popular datasets and investigate the role of low dimensional structure in neural network learning. We find that common natural image datasets indeed have very low intrinsic dimension relative to the high number of pixels in the images. Additionally, we find that low dimensional datasets are easier for neural networks to learn. We validate our findings by carefully-designed experiments to vary the intrinsic dimension of both synthetic and real data and evaluate its impact on sample complexity.

1 Introduction

The idea that real-world data distributions can be described by very few variables underpins machine learning research from manifold learning to dimension reduction (Besold & Spokoiny, 2019; Fodor, 2002). The number of variables needed to describe a data distribution is known as its *intrinsic dimension* (ID). In applications, such as crystallography, computer graphics, and ecology, practitioners depend on data having low intrinsic dimension (Valle & Oganov, 2010; Desbrun et al., 2002; Laughlin, 2014). A variety of deep learning techniques including autoencoders and regularization methods (Gonzalez & Balajewicz, 2018; Zhu et al., 2018) are also motivated by the low-dimensional assumption of data.

It is known that dimensionality plays a strong role in learning function approximations and non-linear class boundaries. The exponential cost of learning in high dimensions is easily captured by the trivial case of sampling a function on a cube; in n dimensions, sampling only the cube vertices would require 2^n measurements. Similar behaviors emerge in learning theory. It is known that learning a manifold requires a number of samples that grows exponentially with the manifold's intrinsic dimension (Narayanan & Mitter, 2010). Similarly, the number of samples needed to learn a well-conditioned decision boundary between two classes is an exponential function of the intrinsic

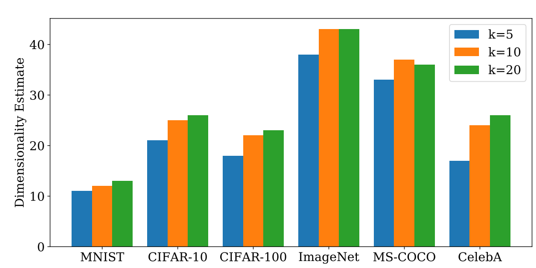

Figure 1: Estimates of the dimensionality of popular datasets obtained using the MLE method with $k = 5, 10, 20$ nearest neighbors (left to right).

Submitted to the Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

dimension of the manifold on which the classes lie (Narayanan & Niyogi, 2009). Furthermore, these learning bounds have no dependence on the ambient dimension in which manifold-structured datasets live.

In light of the exponentially large sample complexity of learning high-dimensional functions, the seemingly low number of samples needed for neural networks to learn image manifolds strongly suggests that image datasets have extremely low-dimensional structure.. Networks learn complex decision boundaries from small amounts of image data, e.g., ImageNet has no more than 1300 images for each of the 1000 classes. At the same time, Generative adversarial networks are able to learn image "manifolds" from merely a few thousand samples.

Despite the established role of low dimensionality in deep learning, little is known about the intrinsic dimension of popular datasets and the impact of dimensionality on the performance of neural networks. We adopt tools from the dimension estimation literature to shed light on dimensionality in settings of interest to the deep learning community. Our contributions are summarized as follows:

- We verify the reliability of intrinsic dimension estimation on high-dimensional data using generative adversarial networks (GANs), a setting where we can upper-bound the intrinsic dimension of generated data a priori by the dimension of the latent noise vector.
- We measure the dimensionality of popular datasets such as MNIST, CIFAR-10, and ImageNet. In our experiments, we find that natural image datasets whose images contain thousands of pixels can, in fact, be described by orders of magnitude fewer variables. For example, we estimate that ImageNet, despite containing 150K pixels per image, only has an intrinsic dimension between 38 and 43; see Figure 1.
- We train classifiers on synthetic and real data of various intrinsic dimension and find that this variable correlates closely with the sample complexity for learning. On the other hand, we find that the dimension of the ambient space of the data has little impact on generalization.

Our results put experimental weight behind the hypothesis that the unintuitively low dimensionality of natural images is being exploited by deep networks, and suggest that a characterization of this structure is an essential building block for a successful theory of deep learning. A brief review of related work can be found in Appendix A.

2 Scalable Estimation of Intrinsic Dimension

Given a set of sample points $\mathcal{P} \subset \mathbb{R}^N$, it is common to assume that $\mathcal{P}$ lies on or near a low-dimensional manifold $\mathcal{M} \subseteq \mathbb{R}^N$ of intrinsic dimension $\dim(\mathcal{M}) = D \ll N$. We implemented a scalable version of the popular Maximum Likelihood Estimator (MLE) of Levina & Bickel (2004); for further information on dimension estimation, see (Kim et al., 2019) and references therein.

2.1 Validation on Synthetic Data

Towards a principled application of ID estimates on images, we begin by validating that MLE methods can generate accurate dimensionality estimates for complex image manifolds. We generate image datasets using generative models for which the intrinsic dimensionality is bounded. We believe such validations are essential to put recent findings in perspective (Gong et al., 2019; Ansuini et al., 2019).

We use the BigGAN variant with 128 latent entries and outputs of size $128 \times 128 \times 3$ trained on the ImageNet dataset (Deng et al., 2009) to generate datasets with various number of images, where we fix most entries of the latent vectors to zero leaving only n free entries to be chosen at random. As we increase the number of free entries, we expect the intrinsic dimension to increase but not to exceed n; see Appendix C.1 for further discussions.

In particular, we create several synthetic datasets of varying intrinsic dimensionality using the ImageNet class, `basenji`, and check if the estimates match our expectation. As seen in Figure 5, we observe increasing diversity with increasing intrinsic dimension. In Figure 6, we show convergence of the MLE estimate on `basenji` data with dimension bounded above by 10. We also observe that the estimates can be sensitive to the choice of k as discussed in prior works; see Appendix C.2 for additional GAN classes. In addition, we evaluate the accuracy of averaging a subset of local MLE estimates for large-scale datasets like ImageNet; see Appendix C.3 for details.

Dataset	MNIST	CIFAR-10	CIFAR-100	ImageNet	MS-COCO	CelebA
MLE (k=5)	11	21	18	38	33	17
MLE (k=10)	12	25	22	43	37	24
MLE (k=20)	13	26	23	43	36	26
SOTA Accuracy	99.84	99.37	93.51	88.5	-	-

Table 1: The MLE estimates for practical image datasets, and the state-of-the-art test-set image classification accuracy (for classification problems only) for these datasets.

2.2 The Intrinsic Dimension of Popular Datasets

In this section, we measure the intrinsic dimensions of a number of popular datasets including MNIST (Deng, 2012), CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), ImageNet (Deng et al., 2009), MS-COCO (Lin et al., 2014), and CelebA (Liu et al., 2015). Using three different parameter settings for MLE, we find that the ID is indeed much smaller than the number of pixels; see Table 2.2. Notice that the rank order of datasets by dimension does not depend on the choice of k. A comparison of state-of-the-art classification accuracy on each respective dataset[1] with the dimension estimates suggests a negative correlation between the intrinsic dimension and test accuracy. In the next section, we take a closer look at this phenomenon through a series of dedicated experiments.

3 Intrinsic Dimension and Generalization

Narayanan & Mitter (2010) have establish that learning a manifold requires a number of samples that grows exponentially with the manifold's intrinsic dimension, but the required number of samples is independent of the extrinsic dimension. We leverage dimension estimation tools to empirically verify these theoretical findings by a family of binary classification problems defined on both synthetic and real datasets of varying ID. We then train classifiers on these datasets and measure test accuracy. In these experiments, we find that classification problems on data of lower intrinsic-dimensionality are easier to solve.

3.1 Synthetic data: Sample complexity depends on intrinsic (not extrinsic) dimensionality

The synthetic GAN data generation technique described in Section B provides a unique opportunity to test the relationship between generalization and intrinsic/extrinsic dimensionality on images. By creating datasets with controlled intrinsic dimensionality, we may compare their *sample complexity*, that is the number of samples required to obtain a given level of test error. Specifically we test the following two hypotheses (1) data of lower intrinsic dimensionality has lower sample complexity than that of higher intrinsic dimensionality and (2) extrinsic dimensionality is irrelevant for sample complexity.

To investigate hypothesis (1), we create four synthetic datasets of varying intrinsic dimensionality: $16, 32, 64, 128$, *fixed* extrinsic dimensionality: $3 \times 128 \times 128$, and two classes: `basenji` and `beagle`. For each dataset we fix a test set of size $N = 1700$. For all experiments, we use the `ResNet-18` (width=64) architecture (He et al., 2016). We then train models until they fit their entire training set with increasing amounts of training samples and measure the test error. We show these results in Figure 2. Observing the varying rates of growth, we see that data of higher intrinsic dimension requires more samples to achieve a given test error.

For hypothesis (2), we carry out the same experiment with the roles of intrinsic and extrinsic dimension switched. We create four synthetic datasets of varying *extrinsic* dimensionality by resizing the images with nearest-neighbor interpolation. Specifically we create 6 datasets of square, 3-channel images of sizes $16, 32, 64, 128, 256$, *fixed* intrinsic dimensionality of size 128, and all other experimental details the same. We show these results in Figure 3. Observing the lack of variable growth rates, we see that extrinsic dimension has little to no effect on sample complexity.

We conclude by noting that, to the best of our knowledge, this is the first such experimental result to demonstrate that *intrinsic but not extrinsic dimensionality matters for the generalization of deep networks*.

[1] Values from https://paperswithcode.com/task/image-classification.

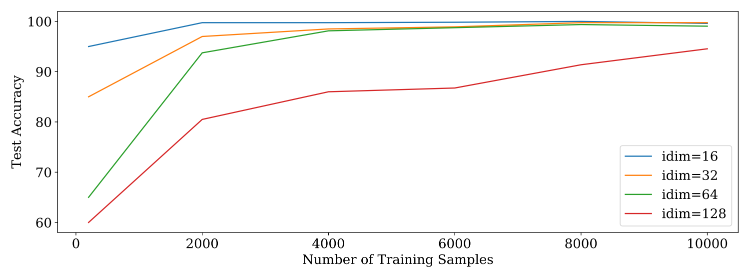

Figure 2: Sample complexity of synthetic datasets of varying intrinsic dimensionality.

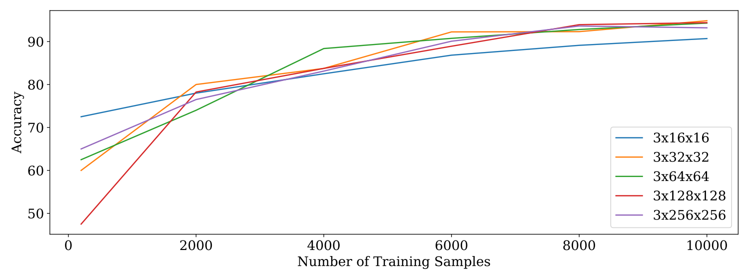

Figure 3: Sample complexity of synthetic datasets of varying extrinsic dimensionality.

3.2 Real data: Adding noise changes dimensionality and affects generalization

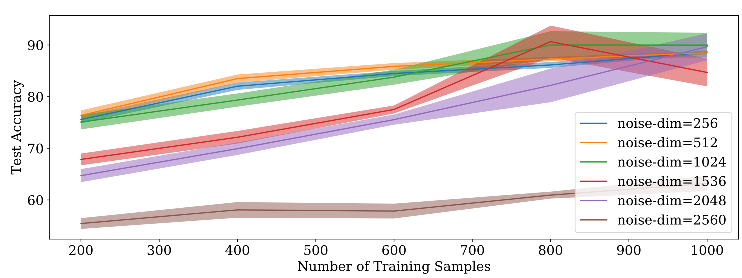

Figure 4: Sample complexity of noisy datasets.

For real data, we evaluate the impact of ID on generalization by: (1) increasing the ID of a dataset by varying amount of additive noise to each image and comparing the test accuracy; (2) comparing the test accuracy of different datasets with different IDs. We defer (2) to Appendix D.

By adding noise to the images of a *real* dataset, we are leveraging the fact that uniformly sampled noise in $[0, 1]^d$ has dimension d. We thus add independent noise, drawn uniformly from a fixed randomly oriented d-dimensional unit hypercube embedded in pixel space, to each sample in a dataset. This procedure ensures that the dataset has dimension at least d. Since we have shown these datasets have low IDs, this procedure specifically increases ID in most cases. We note that estimation error may occur when there is an insufficient number of samples required to achieve a proper dimension estimate. Since the variation in images in a dataset may still be dominated by non-noise directions, we expect to underestimate the new increased dimensions of these noised datasets.

Starting with CIFAR-10 data, we add noise of varying dimensions, where we replace pixels at random in the image. We only add noise to an image once to keep the augmented dataset the same size as the original. We use the following noise dimensionalities: $256, 512, 1024, 2048, 2560$. The new noised data respectively obtains the following MLE dimensionality estimates: $19.7, 30.9, 57.1, 77.8, 110.0, 136.1$. We see that intrinsic dimension increases with increasing noise dimensionality, but dimensionality does not saturate to the maximum true dimension, likely due to a poverty of samples. We show results on these noisy CIFAR-10 datasets in Figure 4. We observe sample complexity largely in the same order as intrinsic dimension.

References

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 6111–6122, 2019.

Franz Besold and Vladimir Spokoiny. Adaptive manifold clustering. *arXiv preprint arXiv:1912.04869*, 2019.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Gunnar Carlsson, Tigran Ishkhanov, Vin de Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1):1–12, 2008.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Mathieu Desbrun, Mark Meyer, and Pierre Alliez. Intrinsic Parameterizations of Surface Meshes. *Computer Graphics Forum*, 2002.

David L. Donoho and Carrie Grimes. Image manifolds which are isometric to euclidean space. *Journal of Mathematical Imaging and Vision*, 23(1):5–24, 2005.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Imola K Fodor. A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Lab., CA (US), 2002.

Sixue Gong, Vishnu Naresh Boddeti, and Anil K Jain. On the intrinsic dimensionality of image representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3987–3996, 2019.

Francisco J Gonzalez and Maciej Balajewicz. Deep convolutional recurrent autoencoders for learning low-dimensional feature dynamics of fluid systems. *arXiv preprint arXiv:1808.01346*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

W. Ronny Huang, Zeyad Emam, Micah Goldblum, Liam Fowl, Justin K. Terry, Furong Huang, and Tom Goldstein. Understanding generalization through visualizations, 2019.

Jisu Kim, Alessandro Rinaldo, and Larry Wasserman. Minimax Rates for Estimating the Dimension of a Manifold. *Journal of Computational Geometry*, 10(1), 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. University of Toronto, 2009. Master's thesis.

Daniel C. Laughlin. The intrinsic dimensionality of plant traits and its relevance to community assembly. *Journal of Ecology*, 102(1):186–193, 2014.

Ann B. Lee, Kim S. Pedersen, and David Mumford. The nonlinear statistics of high-contrast patches in natural images. *International Journal of Computer Vision*, 54(1):83–103, 2003.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, pp. 777–784, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In *Advances in neural information processing systems*, pp. 1786–1794, 2010.

Hariharan Narayanan and Partha Niyogi. On the sample complexity of learning smooth cuts on a manifold. In *COLT*, 2009.

Bruno A Olshausen and David J Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333–339, 1996.

Gabriel Peyré. Manifold models for signals and images. *Computer Vision and Image Understanding*, 113(2): 249 – 260, 2009.

Daniel L Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, 1994.

Mario Valle and Artem R. Oganov. Crystal fingerprint space – a novel paradigm for studying crystal-structure sets. *Acta Crystallographica Section A*, 66(5):507–517, 2010.

Wei Zhu, Qiang Qiu, Jiaji Huang, Robert Calderbank, Guillermo Sapiro, and Ingrid Daubechies. LDMNet: Low dimensional manifold regularized neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

A Related Work

Early work on efficient image representations emphasized the importance of natural image statistics (Ruderman, 1994). It is widely believed that the combination of natural scenes and sensor properties yields very sparse and concentrated image distributions, as has been supported by several empirical studies on image patches (Lee et al., 2003; Donoho & Grimes, 2005; Carlsson et al., 2008). This observation motivated lots of work on efficient coding (Olshausen & Field, 1996) and served as a prior in computer vision (Peyré, 2009). Other work has focused on algorithms for verifying the manifold hypothesis (Fefferman et al., 2016).

The generalization literature seeks to understand why some models generalize better from training data to test data than others. One line of work suggests that the loss landscape geometry explains why neural networks generalize well (Huang et al., 2019). Other generalization work predicts that data with low dimension, along with other properties which do not include extrinsic dimension, characterize the generalization difficulty of classification problems (Narayanan & Niyogi, 2009). In the context of deep learning, Gong et al. (2019) found that neural network features are low-dimensional. Ansuini et al. (2019) further found that the intrinsic dimension of features decreases in late layers of neural networks and observed interesting trends in the dimension of features in early layers. Zhu et al. (2018) recently proposed a regularizer derived from the intrinsic dimension of images augmented with their corresponding feature vectors.

B Validating Dimension Estimation with Synthetic Data

Dimensionality estimates are often applied on “simple” manifolds or toy datasets where the dimensionality is known, and so the accuracy of the methods can be validated. Image manifolds, by contrast, are highly complex, may contain many symmetries and modes, and are of unknown dimension. In principle, there is no reason why MLE-based dimensionality estimates cannot be applied to image datasets. However, because we lack knowledge of the exact dimensionality of image datasets, we cannot directly verify that MLE-based dimensionality estimates scale up to the complexity of image structures.

There is an inherent uncertainty in estimating the ID of a given dataset. First, we cannot be sure if the dataset actually resembles a sampling of points on or near a manifold. Second, there are typically no guarantees that the sampling satisfies the conditions assumed by the ID estimators we are using.

Towards a principled application of ID estimates to the learning context, we begin by validating that MLE methods can generate accurate dimensionality estimates for complex image structures. We do this by generating image datasets using generative models for which the intrinsic dimensionality is known. We believe such validations are essential to put recent findings in perspective (Gong et al., 2019; Ansuini et al., 2019).

GAN Images. We use the BigGAN variant with 128 latent entries and outputs of size $128 \times 128 \times 3$ trained on the ImageNet dataset (Deng et al., 2009). Using this GAN, we generate datasets with a varying number of images, where we fix most entries of the latent vectors to zero leaving only n free entries to be chosen at random. As we increase the number of free entries, we expect the intrinsic dimension to increase, and the output can only be at most n-dimensional; see Section C.1 for further discussion.

In particular, we create several synthetic datasets of varying intrinsic dimensionality using the ImageNet class, `basenji`, and check if the estimates match our expectation. As seen in Figure 5, we observe increasing diversity with increasing intrinsic dimension. In Figure 6, we show convergence of the MLE estimate on `basenji` data with dimension bounded above by 10. We observe that the estimates can be sensitive to the choice of k as discussed in prior work; see Appendix C.2 for additional GAN classes.

Scaling to large datasets. We develop a practical approach for estimating the ID of large datasets such as ImageNet. In this approach, we randomly select a fraction α of the dataset as anchors. Then, we evaluate the MLE estimate using only the anchor points, where nearest-neighbors are computed over the entire dataset. Note that, when anchors are chosen randomly, this acceleration has no impact on the expected value of the result. See Appendix C.3 for an evaluation of this approach.

As with simpler manifolds, we find that there is some variation in the estimate with the choice of k, with $k = 10$ consistently yielding good estimates on this dataset. Despite this variation, it appears that MLE is capable of estimating the dimensionality of this image-structured data within a reasonable margin of error.

C Validation of ID Estimates

In this section, we present additional discussion results and discussion relevant to the ID estimation and related validation experiments in Section B.

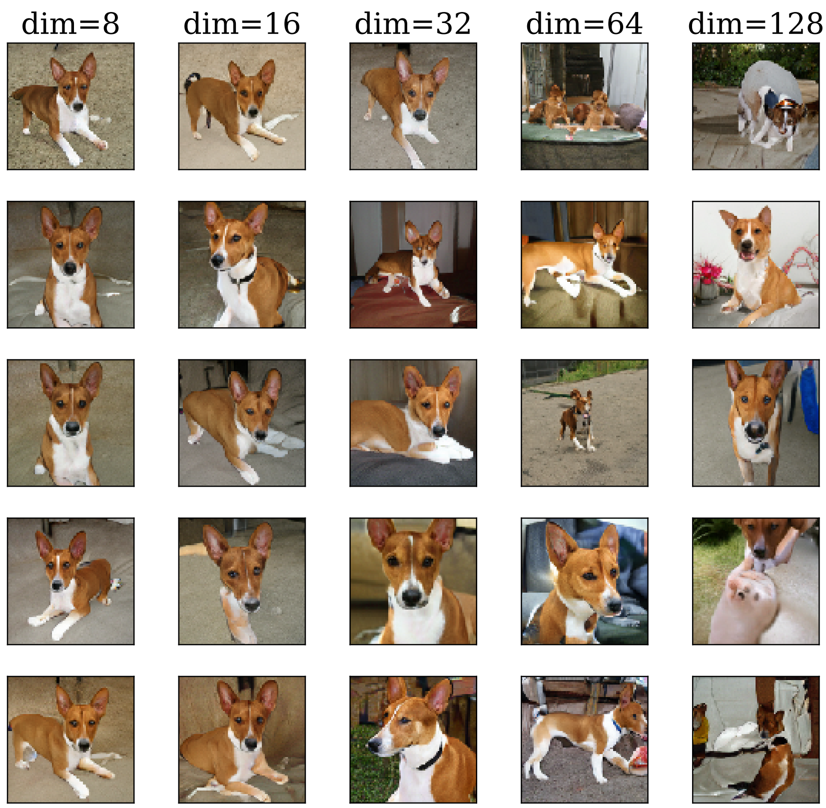

Figure 5: Visualization of `basenji` GAN samples of varying intrinsic dimension.

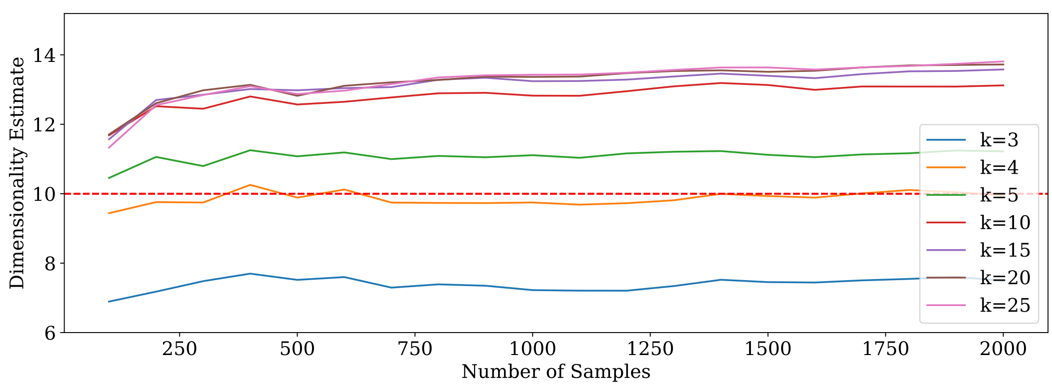

Figure 6: Validation of MLE estimate on synthetic `basenji` data with $n = 10$ free entries. We observe the estimates to converge around the expected dimensionality of 10.

C.1 GAN properties

We devise a method for validating ID measurements in a controlled setting using images generated by GANs. To justify this method, we first note that the image of $\mathbb{R}^d$ under a locally Lipschitz function can be a manifold with dimension at most d. Then, consider that the BigGAN generator, a convolutional neural network with ReLU activations, is a function with this property (Brock et al., 2018).

Specifically, BigGAN can be written as a composition of linear functions, translations, and ReLU activation functions. Individually, these operations do not increase dimension, and by a composition property, their composition cannot increase dimensionality either. The more general fact that the image of $\mathbb{R}^d$ under a locally Lipschitz function can be a manifold with dimension at most d follows from Sard's theorem.

C.2 Convergence for More GAN Classes

We include additional results on the estimation of ID for synthetic GAN images from various ImageNet classes with $n = 10$ free entries. As observed earlier in Section B, the MLE estimates are sensitive to the choice of k, where we expect the ID to be close to 10 given the way we sample the latent vectors to use for the GAN. We note that for a number of classes, all choices of k we considered seem to underestimate the ID.

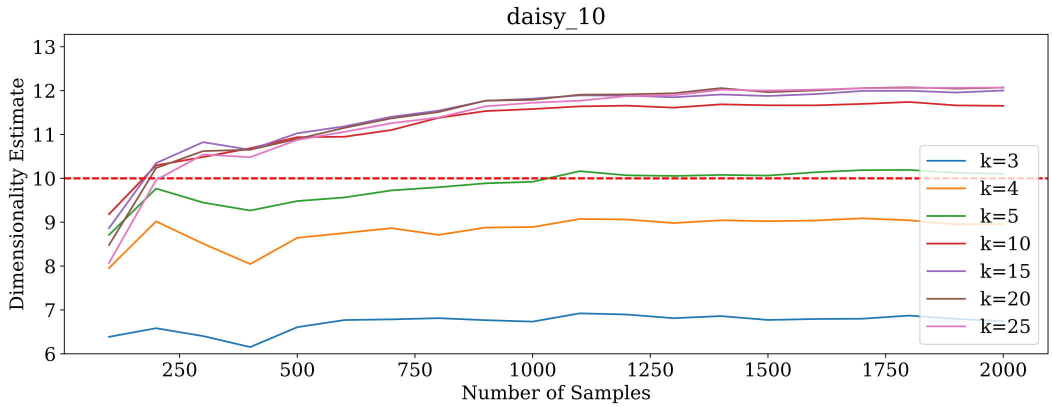

Figure 7: Validation of MLE estimates on synthetic `daisy` data with 10 free entries.

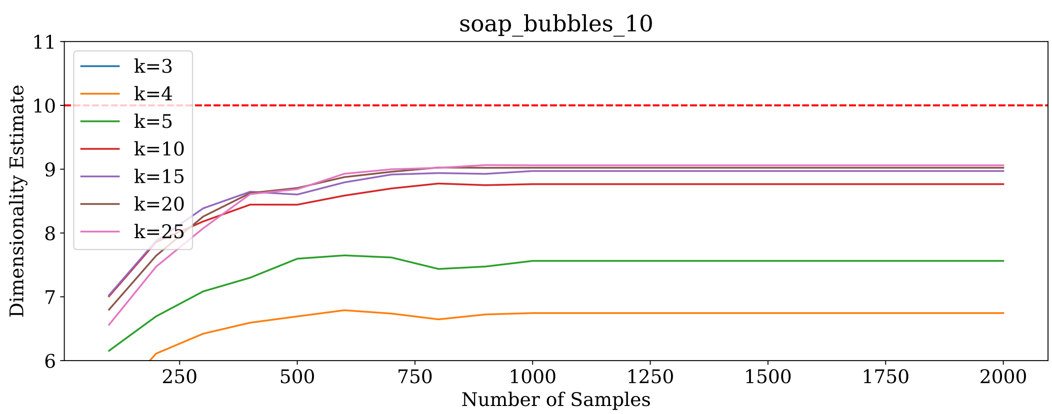

Figure 8: Validation of MLE estimate on synthetic `soap-bubbles` data with 10 free entries.

C.3 Subsampling for Large Datasets

In Figure 10 we validate the anchor approximation on `basenji` data of dimension 10 for varying anchor ratio α. Then, in Figure 11 we validate the anchor approximation on `tree-frog` data of dimension 32 for varying k while fixing the anchor ratio at $\alpha = 0.001$.

D Real Data: Intrinsic dimensionality matters for generalization

We examine the sample complexity of binary classification tasks from four common image datasets: MNIST, SVHN, CIFAR-10, and ImageNet. This case differs from the synthetic case in that we have no control over each dataset's intrinsic dimension. Instead, we estimate it via the MLE method discussed in Section 2. To account for variable difficulty of classes, we randomly sample 5 class pairs from each dataset and run the previously described sample complexity experiment. Note that these subsets differ from those used in Table 2.2, where the estimates are taken from the entire dataset and across all classes.

On these sampled subsets, we find the following mean MLE estimates ($k = 3$): MNIST $\rightarrow 7.5 \pm 0.2$, SVHN $\rightarrow 8.5 \pm 0.1$, CIFAR-10 $\rightarrow 11.4 \pm 0.2$, ImageNet $\rightarrow 15.4 \pm 0.8$. We note that these estimates are consistent with expectation, e.g. MNIST is qualitatively less complex then SVHN and CIFAR-10.

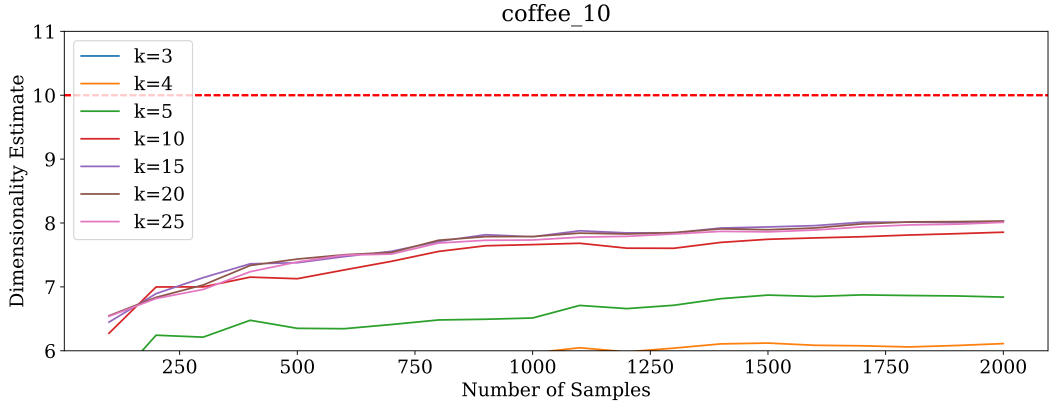

Figure 9: Validation of MLE estimate on synthetic `coffee` data with 10 free entries. Note that the estimates do not converge around the upper bound of 10, which suggests that data generated from this class is not of full dimension.

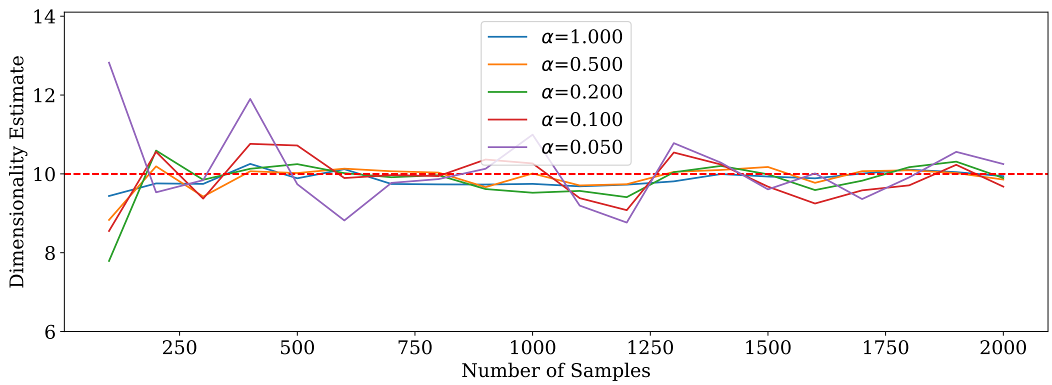

Figure 10: Validation of anchor approximation on `basenji` with 10 free entries.

We conduct the same sample complexity experiment as the previous section on the datasets. Because these datasets are ordinarily of varying extrinsic dimensionality, we resize all to size $32 \times 32 \times 3$ (before applying MLE). We report results in Figure 12, where we overall observe trends ordered by intrinsic dimensionality estimate. These results are consistent with expectation of the relative hardness of each dataset. However, there are some notable differences from the synthetic case. Several unexpected cross-over points exist in the low-sample regime, and the gap between SVHN and CIFAR-10 is smaller than one may expect based on their estimated intrinsic dimension.

From these observations we conclude that intrinsic dimensionality is indeed relevant to generalization on real data, but it is not the only feature of data that influences sample complexity.

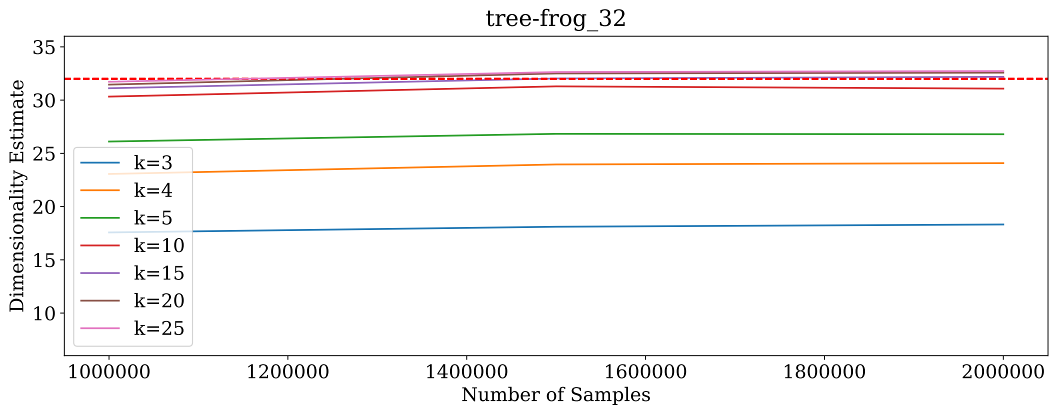

Figure 11: Validation of anchor approximation on `tree-frog` with $n = 32$ free entries.

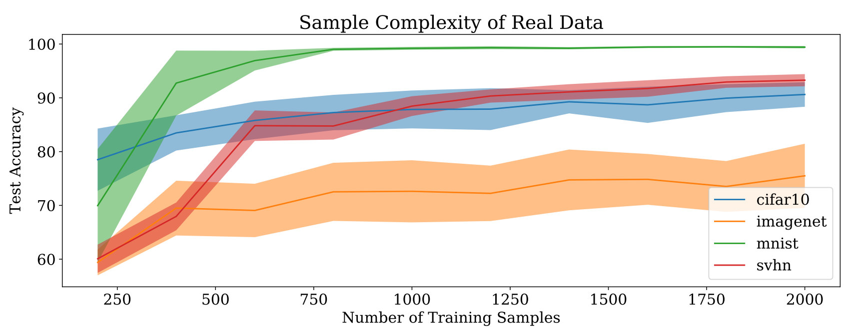

Figure 12: Sample complexity of real datasets.