The Effect of Manifold Entanglement and Intrinsic Dimensionality on Learning

Daniel Kienitz, Ekaterina Komendantskaya, Michael Lones

Heriot-Watt University, Edinburgh, UK
{dk50, e.komendantskaya, m.lones}@hw.ac.uk

Abstract

We empirically investigate the effect of class manifold entanglement and the intrinsic and extrinsic dimensionality of the data distribution on the sample complexity of supervised classification with deep ReLU networks. We separate the effect of entanglement and intrinsic dimensionality and show statistically for artificial and real-world image datasets that the intrinsic dimensionality and the entanglement have an interdependent effect on the sample complexity. Low levels of entanglement lead to low increases of the sample complexity when the intrinsic dimensionality is increased, while for high levels of entanglement the impact of the intrinsic dimensionality increases as well. Further, we show that in general the sample complexity is primarily due to the entanglement and only secondarily due to the intrinsic dimensionality of the data distribution.

Introduction

It is a common assumption that distributions of natural data, such as images, concentrate near or lie on low-dimensional manifolds embedded in high-dimensional ambient spaces (Goodfellow, Bengio, and Courville 2016). The dimension of this manifold is the *intrinsic dimensionality* of the distribution and the dimension of the ambient space is the *extrinsic dimensionality*. It has been shown theoretically that the sample complexity of empirical risk minimization depends on the curvature of the data manifold and the decision boundary, and on the number of intrinsic dimensions, but not on the number of extrinsic dimensions (Narayanan and Niyogi 2009; Narayanan and Mitter 2010). Recently, Pope et al. (Pope et al. 2021) provided empirical evidence that real-world image distributions indeed have low intrinsic dimensionality and that the sample complexity for deep classifiers is positively correlated with the intrinsic and almost independent of the extrinsic dimensionality.

The goal of this work is to further study the effects on the sample complexity of deep classifiers, however, this time under consideration of the *entanglement* of the class manifolds (i.e. the curvature of the decision boundary). Intuitively, the entanglement can be defined as the number of connected hyperplanes that are necessary to perfectly separate the classes. To the best of the authors' knowledge this is the first empirical study that factors in the entanglement when studying the sample complexity and systematically compares the effects of intrinsic dimensionality and entanglement. We define the sample complexity $\varsigma := \Sigma + \mathcal{I} + \mathcal{E}$ as a result of the entanglement Σ, the intrinsic dimensionality $\mathcal{I}$ and the extrinsic dimensionality $\mathcal{E}$ and estimate via regression analysis which of these properties is statistically the most influential one for supervised classification with deep ReLU (Fukushima 1969; Fukushima and Miyake 1982; Glorot, Bordes, and Bengio 2011) networks. We show that the entanglement has a significantly larger effect on the sample complexity than the intrinsic dimensionality and, more importantly, observe an interdependence between these two factors. For low levels of entanglement increasing the intrinsic dimensionality results in equally low increases of the sample complexity while for high levels of entanglement increases in the intrinsic dimensionality lead to larger increases of the sample complexity. In other words, the effect of the intrinsic dimensionality on the sample complexity depends on the given distribution's level of entanglement. Thus, intrinsic dimensionality and entanglement cannot be considered independently when studying the sample complexity but have to be considered jointly.

Our results do not contradict but complement the findings of Pope et al. (Pope et al. 2021). Pope et al.'s investigation of the intrinsic dimensionality's impact on the sample complexity is limited to several complex datasets: ImageNet (Deng et al. 2009), CIFAR-10 (Krizhevsky, Hinton et al. 2009) and FONTS (Stutz, Hein, and Schiele 2019). Thus, their analysis only considers distributions with constant and high levels of entanglement. In our study, on the other hand, we regard the entanglement as another variable that influences the sample complexity and include it in our analysis as well.

The work is structured as follows. In Sections and we describe the notation and related work. Section introduces two simple measures for the entanglement. In Section we demonstrate the aforementioned results first for artificial datasets and then in Section for real-world image benchmarks.

Notation

Throughout this work we consider $l \in \mathbb{N}^+$ samples $x \in \mathbb{R}^{1\times\mathcal{E}}$ arranged in the matrix $X \in \mathbb{R}^{l\times\mathcal{E}}$ with labels $y \in$

$\{0, 1\}$. We assume that those samples concentrate near manifolds $M_{\text{samples}}^{(y=0)} \subset M_{\text{data}}^{(y=0)}$ and $M_{\text{samples}}^{(y=1)} \subset M_{\text{data}}^{(y=1)}$, where $M_{\text{data}}^{(y=0)}$ and $M_{\text{data}}^{(y=1)}$ support the entire data distribution $p(x_{\text{data}})$. A manifold M is a topological space that is locally homeomorphic to a Euclidean space of dimension $\mathcal{I}$, so for every $x \in M$ there exits an open set U, $x \in U \subset M$, that is homeomorphic to an open set $V \subset \mathbb{R}^{\mathcal{I}}$ with homeomorphism $\phi_x : U \to V$. As such, the intrinsic dimensionality can also be described as the dimensionality of the basis that spans the tangent spaces $T_x M$ at points $x \in M$. The dimensionality $\mathcal{I}$ of the aforementioned Euclidean space is the intrinsic dimensionality of the manifold while $\mathcal{E}$ is the dimension of the manifold's ambient space, i.e. the extrinsic dimensionality. For natural images, for example, the extrinsic dimensionality is the number of pixels and colour channels, while the intrinsic dimensions denote the distribution's factors of variation, i.e. those changes that do not alter the semantics of a particular sample. These changes depend on the considered distribution and can for example include rigid transformations, changes in illumination or other changes in the appearance of the objects.

Throughout this work we consider a binary classifier $f : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}_+^2$ which is either a support vector machine with a linear kernel (Boser, Guyon, and Vapnik 1992), a fully-connected or a convolutional neural network. The neural networks have ReLU activations and are trained with the Adam optimizer (Kingma and Ba 2015).

Related Work

In the setting of statistical learning theory (Vapnik 1992) the goal is to find a classifier f that minimizes the *risk* $R(f) \coloneqq \mathbb{E}_{x \in p(x_{\text{data}})}[\mathcal{L}(y, f(x))]$, where $\mathcal{L}$ is a suitable loss function. The *Bayes-classifier* f_{Bayes} is defined as the classifier with the minimum possible risk which parametrizes the conditional distribution $p(y|x_{\text{data}})$. Since $p(y|x_{\text{data}})$ is generally unknown, the goal of learning is to find f that approximates f_{Bayes}. The sample complexity of a hypothesis class containing f is the number of train samples necessary to ensure a probably approximately correct (PAC) solution so a solution such that $|R(f) - R(f_{\text{Bayes}})| < \epsilon$ with probability $1 - \delta$ for $\epsilon, \delta \in \mathbb{R}$.

Narayanan et al. (Narayanan and Niyogi 2009) studied the sample complexity of empirical risk minimization for binary classification from a theoretical point of view. They prove bounds on the sample complexity that depend on the curvature of the data manifold M_{data} on which the data distribution $p_{\text{data}}(x)$ is supported, the curvature of the decision boundary separating $M_{\text{data}}^{(y=0)}$ and $M_{\text{data}}^{(y=1)}$ and the intrinsic dimensionality of M_{data}. Additionally, they show that the extrinsic dimensionality does not have an influence on the sample complexity. Recently, Pope et al. (Pope et al. 2021) confirmed some of these findings for deep classifiers by showing empirically that the sample complexity is well correlated with the intrinsic dimensionality of modern image benchmarks and almost independent of the extrinsic dimensionality. Ansuini et al. (Ansuini et al. 2019) on the other hand studied the intrinsic dimensionality of the data manifold as it is propagated through the network's layers. They find a characteristic increase followed by a progressive decrease of the intrinsic dimensionality and that the intrinsic dimensionality in the last layer is negatively correlated with the generalization error. Brahma et al. (Brahma, Wu, and She 2015) studied the ability of deep belief networks to disentangle and linearise manifolds. They showed that deep architectures progressively linearise and disentangle manifolds and that the presence of extrinsic dimensions that are not predictive of the label can hinder their ability to do so.

Zhang et al. (Zhang et al. 2016), show empirically that neural networks, despite having perfect sample expressivity, generalize well which complicates their analysis by tools from learning theory like the VC-dimension (Harvey, Liaw, and Mehrabian 2017). From a theoretical perspective the generalization capabilities have been studied by several authors (Bartlett 1998; Allen-Zhu, Li, and Liang 2019). Neyshabur et al. (Neyshabur, Bhojanapalli, and Srebro 2018) and Bartlett et al. (Bartlett, Foster, and Telgarsky 2017) provide bounds based on the spectral norms and Lipschitz constant of the networks. Golowich et al. (Golowich, Rakhlin, and Shamir 2018) bound the Rademacher complexity of networks independently of architectural parameters.

Our Work Our work is orthogonal to the aforementioned works as we study the sample complexity not from a model-perspective but from a data-perspective. Since deep classifiers do not always behave like the predictions made by classical statistical learning theory (e.g., (Zhang et al. 2016; Nagarajan and Kolter 2019)) we are interested, whether classical bounds on the sample complexity of empirical risk minimization based on the distribution's geometry hold for deep classifiers. We are especially interested what influence the entanglement of class manifolds has on classifiers since this problem has not been independently studied despite its obvious importance for learning

Entanglement of Class Manifolds

Entanglement Measures

The entanglement between two manifolds can be defined as the number of connected $(\mathcal{E} - 1)$-dimensional hyperplanes needed to perfectly separate the classes. In a two-dimensional ambient space, for example, this corresponds to the number of connected line segments. If two classes are linearly separable, only a single hyperplane is required. Perfect separation is, by definition, given by the Bayes classifier f_{Bayes}. Thus, its decision boundary provides the measure of the entanglement between the two classes. Since f_{Bayes} is unknown, we approximate it with the classifier f. This approximation is a *lower bound* of the true entanglement between classes. Since the available samples $X^{l \times \mathcal{E}}$ are in reality only a small subset of the data distribution $p_{\text{data}}(x)$, they might not be an accurate representation of the topology of the data distribution. If $p(x_{\text{data}})$ is not uniform over M_{data} then, in the worst case, there could be two easily separable modes while the low-density regions are highly entangled. Then, our samples are dominated by the ones coming from the high-density regions and our estimation of the entanglement via investigation of the decision boundary f_d of f will underestimate the true entanglement.

Knowing the actual number of connected line segments necessary to separate the classes implies the availability of a perfect classifier. Thus, computing the absolute level of entanglement for real-world distributions is, from a learning perspective, just as difficult as finding this perfect classifier. In this study, however, we do not require the absolute values of entanglement but only the relative levels. In other words, an ordinal measure that allows to rank different distributions and their subsets according to their entanglement is sufficient for our study. We use the two methods described below.

Linear Support Vector Classifier (LSVC) If f is a support vector classifier with a linear kernel, its accuracy can be used as a measure for the entanglement between two manifolds when compared for different distributions. The poorer the approximation of a decision boundary by a single hyperplane gets, the worse the LSVC's accuracy is if the classes have equal number of samples. Using the LSVC's accuracy this way, we can interpret it as a simple global measure for the entanglement of two manifolds.

Spectrum of the Decision Function's Hessian For the second measure, we consider a neural network classifier f with decision function f_d, where $f_d(\bar{x}) = 0$ for all $\bar{x}$. Assuming a square approximation of the decision function, the second-order Taylor approximation of f_d around $\bar{x}$ yields

$$T_{f_d}(x) = f_d(\bar{x}) + (x - \bar{x})^T \mathcal{J}_{f_d}(\bar{x}) + \frac{1}{2!}(x - \bar{x})^T \mathcal{H}_{f_d}(\bar{x})(x - \bar{x}) \quad (1)$$

where $\mathcal{J}_{f_d}(\bar{x})$ is the Jacobian and $\mathcal{H}_{f_d}(\bar{x})$ is the Hessian of f_d evaluated at $\bar{x}$. Determining the curvature of f_d at $\bar{x}$ where $f_d(\bar{x}) = 0$ comes down to investigating the spectrum of the Hessian $\mathcal{H}_{f_d}(\bar{x})$. In contrast to the LSVC's accuracy this measure of entanglement is local. It quantifies how much the decision boundary around an $\bar{x}$ differs from a linear one.

To compute those $\bar{x}$ for which $f_d(\bar{x}) = 0$ we sample two points of different classes, $x^{(y=0)}$ and $x^{(y=1)}$, and solve

$$\bar{x} = wx^{(y=0)} + (1 - w)x^{(y=1)} \quad (2)$$

for $w \in [0, 1]$. This procedure ensures that all points sampled from the decision boundary are from within the convex hull of the data distribution and therefore separate the two supports, $M^{(y=0)}_{\text{samples}}$ and $M^{(y=1)}_{\text{samples}}$, where they are closest.

Entanglement of Common Image Benchmarks

In this section we test the two entanglement measures introduced in the previous section on the real-world image benchmarks MNIST (LeCun et al. 1990), FASHION (Xiao, Rasul, and Vollgraf 2017), SVHN (Netzer et al. 2011) and CIFAR-10. It is common knowledge that these image benchmarks vary significantly in their entanglement. MNIST, for example, can be solved with high accuracy by a linear classifier while SVHN and CIFAR-10 cannot. In this section we measure the entanglement of the aforementioned datasets by choosing a representative binary classification problem consisting of two semantically similar classes. The intuition behind this is that those classes lie closer in pixel space (and possibly also in an arbitrary representation space) and so might also exhibit greater entanglement (in the presence of nuisance perturbations) than classes that are visually very different. Thus, choosing a representative pair of similar classes for each dataset provides an estimate of the entanglement for the entire dataset. For MNIST and SVHN the similar classes are the digits *eight* and *nine*, for FASHION the classes *ankle boot* and *sneaker* and for CIFAR-10 the classes *cat* and *dog*. To measure the entanglement we balance the classes for the LSVC to make comparisons between datasets possible. We randomly sample a certain fraction of the original binary datasets and compute the LSVC's accuracy on those smaller ones as well as on the complete dataset (*Fraction = 1*). In Figure 1 we display the results for the similar class pair. Unsurprisingly, we observe that the perceived difficulty of these image benchmarks is aligned with this entanglement measure. It is, however, noteworthy that we have to remove a significant fraction of samples of the complex benchmarks SVHN and CIFAR-10 before the LSVC's accuracy improves to levels of that for MNIST and FASHION. This means that a significant amount of samples lie near the decision boundary for those chosen classes.

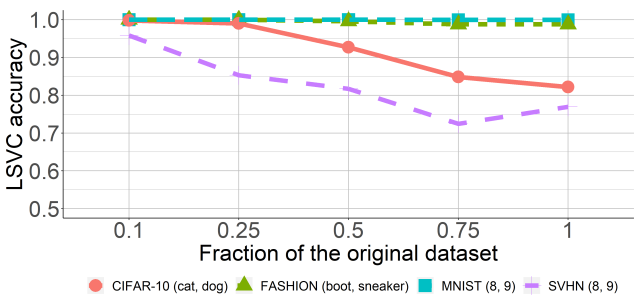


Figure 1: LSVC accuracy for similar classes.

The Hessian entanglement measure gives the same result. We train the neural network classifier f on the class pairs mentioned above and sample 500 points $\{\bar{x}_i\}_{i=1}^{500}$ on its decision boundary f_d for which we compute the Hessian $\mathcal{H}_{f_d}(\bar{x}_i)$. In Figure 2 we display the mean of the ordered singular values of those Hessians. We observe that more complex image datasets, like CIFAR-10 and SVHN, have a higher spectrum and therefore exhibit larger entanglement. Since these results confirm common knowledge and the global LSVC and the local Hessian measure give the same results, we provide only the LSVC's accuracy in our further study.

Intrinsic Dimensionality and Entanglement

When sorted increasingly according to their entanglement the previously used benchmarks exhibit the following order: MNIST $<$ FASHION $<$ SVHN $<$ CIFAR-10 (see Figures 1 and 2). Pope et al. (Pope et al. 2021) report the same order when sorting these benchmarks according to their intrinsic dimensionality. Thus, image datasets with higher intrinsic dimensionality also exhibit higher entanglement.

This observation is noteworthy because in Sections and

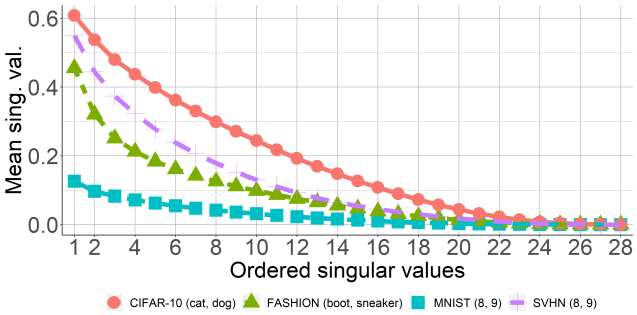


Figure 2: Hessian spectrum for similar classes.

we demonstrate through extensive experimentation on artificial and real-world datasets that the entanglement is the leading contributor to the sample complexity and that the effect of the intrinsic dimensionality depends on the given level of entanglement. Thus, we hypothesize that the reported complexity of these datasets might by primarily due to their entanglement and not their intrinsic dimensionality.

Sample Complexity of Artificial Datasets

In this section we investigate the effect that the entanglement, the intrinsic and the extrinsic dimensionality have on the sample complexity for datasets for which we can control all these parameters independently of each other.

Datasets

Archimedean Spiral Dataset The first artificial dataset consists of one-dimensional Archimedean spirals embedded in a two-dimensional ambient space. In Cartesian coordinates these spirals can be described as

$$A(\Sigma_{\text{Arch}}) = (\Sigma_{\text{Arch}} \cos \Sigma_{\text{Arch}}, \Sigma_{\text{Arch}} \sin \Sigma_{\text{Arch}}) \quad (3)$$

where $\Sigma_{\text{Arch}} \in \mathbb{R}^{\geq 1}$ is the *rotation angle* (see Figure 3a for illustration).

Step-function Dataset The second artificial dataset is directly described by its decision boundary which has the form of a *step-function*. It is defined as

$$S(x) = \lceil x \rceil, \; x \in [1, \Sigma_{\text{Step}}] \quad (4)$$

where $\lceil \cdot \rceil$ denotes the floor function and $\Sigma_{\text{Step}} \in \mathbb{N}^+$ is the maximum value of x (see Figure 3b for illustration).

Changing the Entanglement When we consider two intertwined Archimedean spirals, each generated according to Equation 3 for a common Σ_{Arch}, approximating the decision boundary requires increasingly more linear segments. Therefore, we use the rotation angle Σ_{Arch} as a proxy for the entanglement of the two spirals. For the step-function the maximum value Σ_{Step} of x describes the $(2\Sigma_{\text{Step}} - 1)$ connected line segments that make up the decision boundary, therefore, Σ_{Step} is the proxy for the entanglement.

Increasing the Intrinsic and Extrinsic Dimensionality The original Archimedean spiral dataset is a one-manifold embedded in a two-dimensional ambient space, so it has intrinsic dimensionality $\mathcal{I}_{\text{org}} = 1$ and extrinsic dimensionality $\mathcal{E}_{\text{org}} = 2$. The data separated by the step-functions is a two-manifold embedded in a two-dimensional ambient space as well, so $\mathcal{I}_{\text{org}} = 2$ and $\mathcal{E}_{\text{org}} = 2$. We scale all datasets so that they lie within the unit cube $[0, 1]^{\mathcal{E}}$.

To increase the intrinsic and extrinsic dimensionality of the spiral and the step-function dataset, the original data matrix $X \in \mathbb{R}^{l \times 2}$ generated for some Σ_{Arch} or Σ_{Step} is augmented by a random matrix $I \in \mathcal{U}_{[0,1]}^{l \times \mathcal{I}_{\text{add}}}$, with entries distributed according to a uniform distribution $\mathcal{U}$ over $[0, 1]$, and a zero-matrix $E \in 0^{l \times \mathcal{E}_{\text{add}}}$. $\mathcal{I}_{\text{add}}$ and $\mathcal{E}_{\text{add}}$ are the additional intrinsic and extrinsic dimensions that are added to the base distribution. The augmented data matrix

$$X_a = [X|I|E] \in \mathbb{R}^{l \times (2 + \mathcal{I}_{\text{add}} + \mathcal{E}_{\text{add}})} \quad (5)$$

is matrix-multiplied by a random orthogonal matrix

$$O \in \mathbb{R}^{(2 + \mathcal{I}_{\text{add}} + \mathcal{E}_{\text{add}}) \times (2 + \mathcal{I}_{\text{add}} + \mathcal{E}_{\text{add}})} \quad (6)$$

to remove the previously introduced zeros in the augmented columns. Then, we obtain the projected data matrix

$$X_p = X_a O \in \mathbb{R}^{l \times (2 + \mathcal{I}_{\text{add}} + \mathcal{E}_{\text{add}})} \quad (7)$$

with intrinsic dimensionality $\mathcal{I} = \mathcal{I}_{\text{org}} + \mathcal{I}_{\text{add}}$, extrinsic dimensionality $\mathcal{E} = \mathcal{E}_{\text{org}} + \mathcal{E}_{\text{add}}$ and entanglement Σ_{Arch} or Σ_{Step}, respectively.

Results

Since the spiral and step-function datasets provide an easy way to change the entanglement, intrinsic and extrinsic dimensionality independently of each other, we can estimate the effect that those parameters have on the sample complexity ς. We measure the sample complexity as the number of samples from the train set needed to achieve a certain accuracy on the test set. In other words, we measure the number of samples needed so that the generalization error is below a certain threshold.

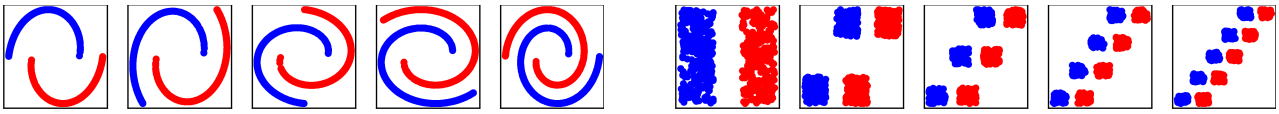
Archimedean Spirals We train a fully-connected neural network on spiral datasets with independently changed $\Sigma_{\text{Arch}} \in [1.0, 1.25, 1.5, 1.75, 2.0]$, $\mathcal{I} \in [1, 2, ..., 11]$ and $\mathcal{E} \in [2, 3, ..., 12]$ and measure the sample complexity ς. Then, we estimate the following three regression models,

$$\varsigma = \alpha \Sigma_{\text{Arch}} + \beta \mathcal{I} + \gamma \mathcal{E} \quad (8)$$

$$\begin{aligned} \varsigma = &\alpha \Sigma_{\text{Arch}} + \beta \mathcal{I} + \gamma \mathcal{E} + \\ &\delta(\Sigma_{\text{Arch}} \cdot \mathcal{I}) + \epsilon(\Sigma_{\text{Arch}} \cdot \mathcal{E}) + \zeta(\mathcal{I} \cdot \mathcal{E}) \end{aligned} \quad (9)$$

$$\varsigma = \beta \mathcal{I} + \gamma \mathcal{E} + \sum_{\sigma \in \Sigma_{\text{Arch}}} \hat{\alpha}^{(\sigma)} [\Sigma_{\text{Arch}}^{(\sigma)}] + \alpha^{(\sigma)} (\mathcal{I} \cdot [\Sigma_{\text{Arch}}^{(\sigma)}]) \quad (10)$$

where $\hat{\alpha}^{(\cdot)}, \alpha^{(\cdot)}, \beta, \gamma, \delta, \epsilon, \zeta \in \mathbb{R}$ are the regression coefficients. $[\Sigma_{\text{Arch}}^{(\cdot)}]$ denotes dummy variables for different entanglement values. The dummy shows the level of entanglement when it is given or zero otherwise. The base case is



(a) $\Sigma_{\text{Arch}} = [1.0, 1.25, 1.5, 1.75, 2.0]$ (b) $\Sigma_{\text{Step}} = [1, 2, 3, 4, 5]$

Figure 3: (a) Archimedean spiral datasets. (b) Step-function datasets.

$[\Sigma_{\text{Arch}}^{(1.0)}]$ and it is therefore omitted from the regression equation.

The first two regression models displayed in Equations 8 and 9 measure the effect of the entanglement, the intrinsic and extrinsic dimensionality independently of each other and with a potential interaction between them. The introduction of the dummy variables in Equation 10 allows us to estimate the intrinsic dimensionality's effect on the sample complexity given a certain level of entanglement. These regression models offer the best trade-off between interpretability and goodness-of-fit. Choosing higher-order polynomials to model the interactions between independent variables might result in a better fit; however, we would sacrifice model interpretability, as well as risk overfitting noise.

The results of these regressions are displayed in Table 1. We observe that in all three cases the entanglement is by a significant margin the most impactful factor on the sample complexity while the extrinsic dimensionality is not statistically relevant. In addition, we can state that the effect of the intrinsic dimension on the sample complexity depends on the distribution's entanglement. While for easily separable datasets ($\Sigma_{\text{Arch}} = [1, 1.25, 1.5]$) increases in the intrinsic dimensionality do not influence the sample complexity significantly, we can observe that for highly entangled datasets ($\Sigma_{\text{Arch}} = [1.75, 2.0]$) the sample complexity positively increases with an increase of the intrinsic dimensionality. In other words, the combination of intrinsic dimensionality and entanglement is empirically the most important one for the difficulty of the learning problem and when judging the sample complexity for a certain classification problem both of these parameters cannot be investigated independently but need to be considered in conjunction.

Step-function Datasets For the step-function datasets we estimate the same regression models as for the Archimedean spirals, so Equations 8 and 9 but with Σ_{Step} instead of Σ_{Arch}. The regression in Equation 10 is estimated for the levels $\Sigma_{\text{Step}} = [1, 2, 3, 4, 5]$ where $[\Sigma_{\text{Step}} = 1]$ is the base case. Again, we train a fully-connected neural network and measure the sample complexity.

In Table 2 we display the findings and can observe that the results are aligned with the ones for the Archimedean spirals. Again, the entanglement is the significantly more important factor for the sample complexity. The previously made observation that the intrinsic dimension's influence depends on the given entanglement is similar for the step-function datasets. In Table 2 we can see that the intrinsic dimensionality positively influences the sample complexity for all levels of entanglement. However, this increase is larger for higher levels of entanglement, so the earlier made observation of an interdependent effect of intrinsic dimensionality and entanglement on the sample complexity remains true. In contrast to the results of the Archimedean spiral datasets, we observe for Equation 10 in Table 2 a statistically significant *negative* effect of the extrinsic dimensionality on the sample complexity. This results is not theoretically

	Eq. 8	Eq. 9	Eq. 10
Σ_{Arch}	167.87^{***}	-32.13^{*}	
	(6.97)	(17.19)	
$\mathcal{I}$	10.06^{***}	-39.83^{***}	0.07
	(0.78)	(3.05)	(0.92)
$\mathcal{E}$	0.53	-1.65	0.53
	(0.78)	(2.97)	(0.41)
$\Sigma_{\text{Arch}} * \mathcal{I}$		32.45^{***}	
		(1.76)	
$\Sigma_{\text{Arch}} * \mathcal{E}$		0.75	
		(1.76)	
$\mathcal{I} * \mathcal{E}$		0.17	
		(0.20)	
$[\Sigma_{\text{Arch}}^{(1.25)}]$			0.30
			(8.85)
$[\Sigma_{\text{Arch}}^{(1.5)}]$			-1.07
			(8.85)
$[\Sigma_{\text{Arch}}^{(1.75)}]$			-19.75^{**}
			(8.85)
$[\Sigma_{\text{Arch}}^{(2.0)}]$			-23.54^{***}
			(8.85)
$\mathcal{I} * [\Sigma_{\text{Arch}}^{(1.25)}]$			0.16
			(1.31)
$\mathcal{I} * [\Sigma_{\text{Arch}}^{(1.5)}]$			1.28
			(1.31)
$\mathcal{I} * [\Sigma_{\text{Arch}}^{(1.75)}]$			15.70^{***}
			(1.31)
$\mathcal{I} * [\Sigma_{\text{Arch}}^{(2.0)}]$			32.79^{***}
			(1.31)
Constant	-254.58^{***}	52.71^{*}	6.04
	(12.92)	(27.76)	(6.89)
Observations	605	605	605
R^2	0.55	0.72	0.88
Adjusted R^2	0.55	0.71	0.87
F Statistic	249.28^{***}	250.96^{***}	421.07^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 1: Regression: Archimedean spirals.

	Eq. 8	Eq. 9	Eq. 10
Σ_{Step}	27.28***	27.08***	
	(0.60)	(1.77)	
$\mathcal{I}$	6.15***	3.27***	2.84***
	(0.27)	(0.83)	(0.53)
$\mathcal{E}$	−1.75***	0.58	−1.75***
	(0.27)	(0.78)	(0.24)
$\Sigma_{\text{Step}} * \mathcal{I}$		0.95***	
		(0.18)	
$\Sigma_{\text{Step}} * \mathcal{E}$		−0.78***	
		(0.18)	
$\mathcal{I} * \mathcal{E}$		0.004	
		(0.08)	
$[\Sigma^{(2)}_{\text{Step}}]$			19.37***
			(5.09)
$[\Sigma^{(3)}_{\text{Step}}]$			55.75***
			(5.09)
$[\Sigma^{(4)}_{\text{Step}}]$			65.80***
			(5.09)
$[\Sigma^{(5)}_{\text{Step}}]$			84.73***
			(5.09)
$\mathcal{I} * [\Sigma^{(2)}_{\text{Step}}]$			4.31***
			(0.75)
$\mathcal{I} * [\Sigma^{(3)}_{\text{Step}}]$			3.26***
			(0.75)
$\mathcal{I} * [\Sigma^{(4)}_{\text{Step}}]$			4.16***
			(0.75)
$\mathcal{I} * [\Sigma^{(5)}_{\text{Step}}]$			4.82***
			(0.75)
Constant	−7.39**	−6.60	29.33***
	(3.15)	(6.80)	(3.97)
Observations	605	605	605
R^2	0.82	0.83	0.86
Adjusted R^2	0.82	0.83	0.85
F Statistic	891.43***	485.05***	352.83***
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 2: Regression: Step-functions.

predicted and since the other findings are aligned with the previous experimental results, we hypothesise that it might be due to the topology of the step-function dataset which is the union of disjoint linear subspaces. An investigation into what causes this effect is left for future work.

Summary

For both artificial datasets we observe that the regressions that take interactions between the entanglement and the intrinsic dimensionality into account fit the observed sample complexities significantly better than the regressions that assume their independence. These results show a statistically significant interaction between these two factors and demonstrate that the effect of the intrinsic dimensionality on the sample complexity is dependent on the given level of entanglement. For datasets that exhibit low levels of entanglement (so those that are (almost) linearly separable), increases in their intrinsic dimensionality have either no or small effects on the sample complexity relative to complex datasets where classes are highly entangled.

Sample complexity of Real-world Datasets

We now expand the analysis from the previous section to real-world image benchmarks.

Datasets

We use the binary classification problems introduced in Section again. Since FASHION is (almost) linearly separable even for large sample sizes, we defer its analysis to the extended on-line version where we show that increases in its intrinsic dimensionality do not appear to cause an increase in the sample complexity. In this Section we only present the results for SVHN and CIFAR-10.

Changing the Entanglement As discussed in Section the number of samples drawn from the data distribution $p(x_{\text{data}})$ can influence the estimation of the entanglement when the density is not uniform over the data manifold M_{data}. Therefore, estimation of the entanglement via a well-trained classifier f only gives a lower bound on it. As a result, without access to $p(x_{\text{data}})$ from which we could sample, we cannot increase but only decrease the entanglement of a given distribution. To decrease the entanglement between two manifolds, the class-boundary points, those samples close to the decision boundary, need to be removed. One way to identify these points is by computing the magnitude of a neural network's gradient $g_i = ||\frac{\partial \mathcal{L}_f}{\partial x_i}||_F$ for all train samples $x_i \in X^{l \times \mathcal{E}}$, where $\mathcal{L}_f$ is the network's loss function and $||\cdot||_F$ denotes the Frobenius-norm. Then, g_i can be used as an estimate of the proximity of point x_i to the boundary. We remove those points with values g_i above the Σ_{Real}-percentile of all gradient norms and replace them with random samples from the class interior, perturbed by Gaussian noise. For $\Sigma_{\text{Real}} = 0.4$ for example, those 60% of points which have the highest gradient norms g_i are removed. $\Sigma_{\text{Real}} = 1$ is the original set. We test this heuristic and can confirm that this approach indeed reduces the entanglement up to some negligible stochastic effects (Figure 4). We note that a significant number of samples need to be removed to observe a meaningful reduction in the entanglement. This is in line with the findings described in Section in which we show that the samples of complex image benchmarks appear to concentrate near the decision boundary.

Increasing the Intrinsic Dimensionality To increase the intrinsic dimensionality of the image datasets we use a similar procedure as in Section . We add uniform noise over $[0, 1]^{\mathcal{I}_{\text{add}}}$ to all available samples. This procedure has been shown to increase the intrinsic dimensionality of real-world image datasets (Pope et al. 2021).

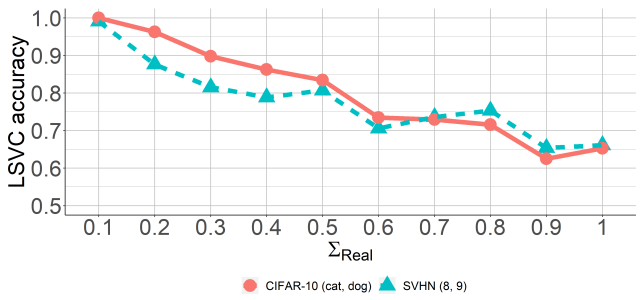


Figure 4: Entanglement of SVHN and CIFAR-10 classes.

Results

We train a convolutional neural network with batch-normalization (Ioffe and Szegedy 2015) on the binary classification tasks for $\Sigma_{\text{Real}} \in [0.1, 0.5, 1.0]$ and $\mathcal{I}_{\text{add}} \in [0, 5, 10, 15, 30, 60, 90, 120, 150]$ where $\mathcal{I}_{\text{add}} = 0$ is the dataset with the original intrinsic dimensionality. Pope et al. (Pope et al. 2021) report an original intrinsic dimensionality for SVHN between 9 and 19 and for CIFAR-10 between 13 and 25, depending on the method. Thus, the ratios between intrinsic and extrinsic dimensionality for the artificial and the real-world datasets are comparable in our work.

We estimate the same regressions as for the artificial datasets (Equations 8, 9 and 10) with the exceptions that we have omitted changing the extrinsic dimensionality as we have have not found it to be statistically significant and that we normalize the sample complexity by dividing it by the number of available samples; $\varsigma_{\text{norm}} = \frac{\varsigma}{l}$ to make comparisons between different class pairs and datasets possible.

The results are displayed in Tables 3 and 4. As for the artificial datasets the entanglement is in general the most significant factor for the (normalized) sample complexity. Importantly, we note again a dependence of the intrinsic dimensionality's impact on the level of entanglement. When enough samples from the class boundaries of both classification tasks are removed, increasing the intrinsic dimensionality does not have a statistically significant effect on the sample complexity any more.

Conclusions

The sample complexity of empirical risk minimization has been studied theoretically and recent empirical work has confirmed that effect of the intrinsic dimensionality on the sample complexity of deep classifiers. In addition, theoretical bounds on the sample complexity of deep classifiers have been proposed (see Section). In this work we take an orthogonal approach to the model-dependent bounds on the sample complexity and provide an extension for the data-dependent study. This is achieved by investigating the effect of the entanglement of class manifolds on the sample complexity. We show for deep ReLU networks that the entanglement is the most important factor for the difficulty of a learning problem and that it has an interdependent effect with the intrinsic dimensionality. Fully-connected and convolutional classifiers exhibit much stronger increases of their sample complexity for higher levels of entanglement, while for low levels the intrinsic dimensionality's effect is smaller.

	Eq. 8	Eq. 9	Eq. 10
Σ_{Real}	0.084^{***}	0.066^{***}	
	(0.008)	(0.010)	
$\mathcal{I}_{\text{add}}$	0.0001^{**}	-0.00004	0.00002
	(0.0001)	(0.0001)	(0.0001)
$\mathcal{I}_{\text{add}} \cdot \Sigma_{\text{Real}}$		0.0003^{**}	
		(0.0001)	
$[\Sigma_{\text{Real}}^{(0.5)}]$			0.014^{*}
			(0.008)
$[\Sigma_{\text{Real}}^{(1.0)}]$			0.058^{***}
			(0.008)
$\mathcal{I}_{\text{add}} \cdot [\Sigma_{\text{Real}}^{(0.5)}]$			0.0001
			(0.0001)
$\mathcal{I}_{\text{add}} \cdot [\Sigma_{\text{Real}}^{(1.0)}]$			0.0003^{***}
			(0.0001)
Constant	0.001	0.011	0.022^{***}
	(0.006)	(0.007)	(0.005)
Observations	27	27	27
R^2	0.832	0.867	0.920
Adjusted R^2	0.818	0.850	0.900
F Statistic	59.582^{***}	50.163^{***}	47.984^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 3: Regression: SVHN-(8, 9).

	Eq. 8	Eq. 9	Eq. 10
Σ_{Real}	0.112^{***}	0.094^{***}	
	(0.011)	(0.016)	
$\mathcal{I}_{\text{add}}$	0.0001	-0.0001	-0.000
	(0.0001)	(0.0001)	(0.0001)
$\mathcal{I}_{\text{add}} \cdot \Sigma_{\text{Real}}$		0.0003	
		(0.0002)	
$[\Sigma_{\text{Real}}^{(0.5)}]$			0.006
			(0.007)
$[\Sigma_{\text{Real}}^{(1.0)}]$			0.082^{***}
			(0.007)
$\mathcal{I}_{\text{add}} \cdot [\Sigma_{\text{Real}}^{(0.5)}]$			0.00004
			(0.0001)
$\mathcal{I}_{\text{add}} \cdot [\Sigma_{\text{Real}}^{(1.0)}]$			0.0003^{***}
			(0.0001)
Constant	-0.020^{**}	-0.011	0.010^{*}
	(0.008)	(0.010)	(0.005)
Observations	27	27	27
R^2	0.808	0.827	0.962
Adjusted R^2	0.792	0.805	0.952
F Statistic	50.587^{***}	36.666^{***}	105.066^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4: Regression: CIFAR-10-(cat, dog).

References

Allen-Zhu, Z.; Li, Y.; and Liang, Y. 2019. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 6158–6169.

Ansuini, A.; Laio, A.; Macke, J. H.; and Zoccolan, D. 2019. Intrinsic dimension of data representations in deep neural networks. In *NeurIPS*.

Bartlett, P. L. 1998. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2): 525–536.

Bartlett, P. L.; Foster, D. J.; and Telgarsky, M. J. 2017. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30: 6240–6249.

Boser, B. E.; Guyon, I. M.; and Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152.

Brahma, P. P.; Wu, D.; and She, Y. 2015. Why deep learning works: A manifold disentanglement perspective. *IEEE transactions on neural networks and learning systems*, 27(10): 1997–2008.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Fukushima, K. 1969. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4): 322–333.

Fukushima, K.; and Miyake, S. 1982. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, 267–285. Springer.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323. JMLR Workshop and Conference Proceedings.

Golowich, N.; Rakhlin, A.; and Shamir, O. 2018. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, 297–299. PMLR.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Harvey, N.; Liaw, C.; and Mehrabian, A. 2017. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Conference on learning theory*, 1064–1068. PMLR.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456. PMLR.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

LeCun, Y.; Boser, B. E.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W. E.; and Jackel, L. D. 1990. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, 396–404.

Nagarajan, V.; and Kolter, J. Z. 2019. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32.

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

Narayanan, H.; and Niyogi, P. 2009. On the Sample Complexity of Learning Smooth Cuts on a Manifold. In *COLT*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

Neyshabur, B.; Bhojanapalli, S.; and Srebro, N. 2018. A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks. In *International Conference on Learning Representations*.

Pope, P.; Zhu, C.; Abdelkader, A.; Goldblum, M.; and Goldstein, T. 2021. The Intrinsic Dimension of Images and Its Impact on Learning. *arXiv preprint arXiv:2104.08894*.

Stutz, D.; Hein, M.; and Schiele, B. 2019. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6976–6987.

Vapnik, V. 1992. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, 831–838.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.