CW Complex Hypothesis for Image Data

Yi Wang [*1] **Zhiren Wang** [*2]

Abstract

We examine both the manifold hypothesis (Bengio et al., 2013) and the union of manifold hypothesis (Brown et al., 2023), and argue that, in contrast to these hypotheses, the local intrinsic dimension varies from point to point even in the same connected component. We propose an alternative CW complex hypothesis that image data is distributed in "manifolds with skeletons". We support the hypothesis by visualizing distributions of 2D families of synthetic image data, as well as by introducing a novel indicator function and testing it on natural image datasets. One motivation of our work is to explain why diffusion models have difficulty generating accurate higher dimensional details such as human hands. Under the CW complex hypothesis and with both theoretical and empirical evidences, we provide an interpretation that the mixture of higher and lower dimensional components in data obstructs diffusion models from efficient learning.

1. Introduction

The manifold hypothesis was proposed in (Bengio et al., 2013) and has been a common assumption on distributions of natural datasets in many researches in various areas of machine learning, especially for image datasets. It states that the distribution of data in a large data set, up to small errors, often lies in a lower-dimensional submanifold of the ambient Euclidean space. The dimension of this lower dimensional-manifold is commonly referred to as the intrinsic dimension. Previous research, such as (Schölkopf et al., 1998), (Tenenbaum et al., 2000), (Roweis & Saul., 2000), (Brand, 2002), (Fodor, 2002), (Ozakin & Gray, 2009), (Narayanan & Mitter, 2010), (Besold & Spokoiny, 2019), (Ansuini et al., 2019) among many others, supported this hypothesis for a wide variety of image datasets. A comprehensive work on empirical verification of manifold hypothesis for many commonly-used image datasets was carried out in (Pope et al., 2021). See also (Fefferman et al., 2016) for principled algorithms for verifying the manifold hypothesis. Many works related manifold hypothesis to deep generative models (DGMs), e.g.(Rezende et al., 2020), (Brehmer & Cranmer, 2020), (Mathieu & Nickel, 2020), (Arbel et al., 2021), (Kothari et al., 2021), (Caterini et al., 2021). The literature relevant to this area extends beyond the references above.

The union of manifold hypothesis (Brown et al., 2023) challenged the manifold hypothesis and suggested that the data is distributed on a disjoint union of submanifolds, which may have different dimensions. This new hypothesis is supported by the experiments in that paper, which show that there is indeed variation in intrinsic dimensions across different sample classes in the same dataset. An alternative way of stating the hypothesis from (Brown et al., 2023) is that each connected component of datasets is a manifold equipped with a measure with continuous density.

The current paper examines to what extent the hypothesis that connected components are manifolds is valid. Based on empirical evidences and simple modeling, we find the need of an even less restrictive hypothesis than that of (Brown et al., 2023), so that intrinsic dimension is allowed to vary within each connected component. In the generic setting, we expect a dataset to be supported on a disjoint union of "manifolds with skeletons", a.k.a. CW complexes, where each connected component is a CW complex consisting of finitely many bounded submanifolds glued together along their boundaries. In particular, the distribution of data may show concentration on lower dimensional skeletons.

The contributions of this paper can be summarized as:

1. In Section 2.1 and Appendix A.1, we create synthetic image data of a 2D family of randomly positioned shapes and visualized the distribution by projecting to a 2D plane. The visualizations show concentration along 1D skeletons.

2. In Proposition 2.2, we introduce a new indicator function I_k measuring the variance of local intrinsic dimension across data points, and prove $I_k \approx 1$ under the manifold hypothesis. Experiments show that I_k is indeed around 1 for true manifold data, but much larger for label classes

[*]Equal contribution [1]Department of Mathematics, Johns Hopkins University [2]Department of Mathematics, Pennsylvania State University. Correspondence to: Yi Wang <ywang261@jhu.edu>, Zhiren Wang <zhirenw@psu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

of natural image datasets, suggesting that even connected components are far from satisfying the manifold hypothesis.

3. We outline reasons in Section 3.1 why image data display high variance in local intrinsic dimensions, even between nearby points.

4. We formulate in Section 3.2 the CW complex hypothesis, and prove that it is compatible with deep learning with ReLu activation function.

5.We offer an explanation on why diffusion based generative models have difficulty generating complicated higher dimensional local features, such as hands in portraits. Based on the CW complex hypothesis, we reason in Section 4.2.1 that the score function in diffusion models, which is the true target function for neural networks to learn, along the coordinates that depict extra features, has much larger magnitude near lower dimensional components than near higher dimensional ones, encouraging the neural network to ignore the higher dimensional part. We give mathematical proofs of such differences in magnitude, and design an experiment in Section 4.2.2 on a modified MNIST dataset consisting of a high dimensional component and a lower dimensional one to support our interpretation.

Comparison with past literature The study of the manifold hypothesis mainly focused on evidences based on the success of using manifold-based statistics such as geodesic distances and curvatures, as well as manifold learning methods such as Multi-dimensional Scaling (Kruskal, 1964), IsoMap (Tenenbaum et al., 2000), Local Linear Embedding (Roweis & Saul., 2000), Laplacian EigenMaps (Belkin & Niyogi, 2003), and Hessian EigenMap (Donoho & Grimes, 2003), in recovering data. The challenge from (Brown et al., 2023) focused on the difference between intrinsic dimensions across classes. Our approach in estimating intrinsic dimension in §2 is similar to that of (Pope et al., 2021; Brown et al., 2023). Namely we also apply the Maximal Likelihood Estimator from (Levina & Bickel, 2004; MacKay & Zoubin, 2005). However, our indicator function is novel and we focus on the new observation of local intrinsic dimension variations between individual data points (within the same dataset class). Combining this new insight with the implication of local manifold structure from the manifold learning research list above, we naturally derive the CW complex hypothesis.

A manifold is a structure enjoying two properties: (i) consistency of local dimensions across data points, (ii) regularity of local geometry (the local shape can be approximated by a linear subspace). Our paper challenges property (i) in manifold hypothesis for both (naturally derived) synthetic data and natural datasets. The support to property (ii) in literature largely relied on the success of using manifold-based statistics such as geodesic distances and curvatures, as well as manifold learning methods such as Multi-dimensional Scaling (Kruskal, 1964), IsoMap (Tenenbaum et al., 2000), Local Linear Embedding (Roweis & Saul., 2000), Laplacian EigenMaps (Belkin & Niyogi, 2003), and Hessian EigenMap (Donoho & Grimes, 2003), in recovering data. While updating assumption (i), we intend to keep assumption (ii), and the resulting structures are exactly characterized by CW-complexes. On the other hand, it might be of interest to view validation of (ii) as a future research direction as well, but that is not the main focus of the current paper.

2. Evidences for non-manifold distributions of image data

In this section, we present two evidences that not only the support of image datasets, but also their connected components, fail to be manifolds. Instead, for a fixed connected component X_i, the data distribution has different dimensions at different points $x \in X_i$.

2.1. Images of random geometric objects

We produce a synthetic image dataset of grayscale images of a continuous 2-dimensional family of geometries. We randomly place a rectangle and a disk in the plane. The side lengths and position of the rectangle are linearly parameterized by a single parameter $t \in [0, 1]$. Similarly, the radius and position of the disk are linearly parameterized by a single parameter s. For a pair (t, s), take a photo shot of the resulting geometry in a fixed square window $[0, 1]^2$ and save it as a gray scale image of $r \times r$ pixels. Each pixel represents a $\frac{1}{r} \times \frac{1}{r}$ square, and its assigned grayscale value is the portion of this pixel square covered by the union of both shapes. Both shapes can overlap with each other, or be either entirely or partially out of canvas. We sample many $(t_i, s_i) \in [0, 1]^2$ uniformly and generate a data set $\mathcal{S}$. Each image is stored as a point $x_i \in [0, 1]^{r\times r}$.

Note that the map $\phi : (t_i, s_i) \to x_i$ is a continuous and piecewise smooth map from $[0, 1]^2$ to $[0, 1]^{r\times r}$. In particular, the distribution of $\mathcal{S}$ is at most 2-dimensional and its support is connected. Thus by Marstrand's projection theorem (Marstrand, 1954), for a generic random linear projection $\pi : [0, 1]^{r\times r} \to [0, 1]^2$ the dimension of $\pi(\mathcal{S})$ is the same as that of $\mathcal{S}$. The local geometry of $\mathcal{S}$ near $\phi(t_i, s_i)$ is approximately inside the image vector space of $D\phi(t_i, s_i)$, which has dimension ≤ 2. Generically, π is non-degenerate on this vector space. Thus a random projection of $\mathcal{S}$ can be understood as a geometrically faithful view of $\mathcal{S}$ from random angles. Such 2D visualization is the reason that we choose to use a 2-dimensional family of images.

Experiments with $|\mathcal{S}| = 50000$ and $r = 16$ produced the visualization in Figure 1. The first picture contains sample images from $\mathcal{S}$ and the second depicts a random projection of $\mathcal{S}$ into $\mathbb{R}^2$. The "skeletons" shows that there is significant

concentration of mass along 1-dimensional submanifolds while a large part of the distribution are scattered on 2 dimensional surfaces.

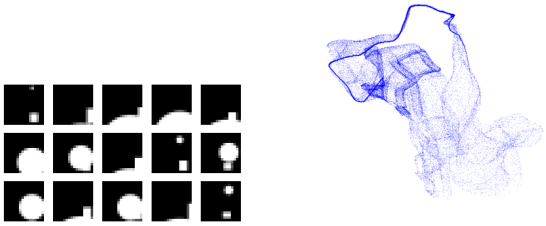

Figure 1. 16x16 random geometric images and their distributions

Coloring each data point x_i according to its estimated local intrinsic dimension $\hat{d}_k(x_i)$ with $k = 100$, which will be defined later in (18), we produce a color enhanced visualization in Figure 2. It shows that the local dimension of all points near the 1 dimensional skeleton are around 1. Indeed, using $\hat{d}_k(x_i) < 1.5$ as cutoff, the 1-dimensional curve (red colored) accounts for 27% of the entire dataset.

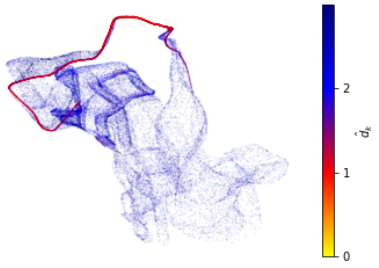

Figure 2. Visualization as in Figure 1, colored based on $\hat{d}_{100}(x_i)$

The skeleton phenomenon in the visualization is not limited to geometric shapes. A similar experiment using MNIST-based shapes can be found in Appendix A.1.

2.2. Criterion on distribution of local intrinsic dimension estimates in manifolds

We now introduce an indicator function to test whether a distribution is supported on a manifold, and prove by experiments that this criterion fails on individual label classes from natural image datasets.

Following the approach from (Brown et al., 2023) and (Pope et al., 2021), we use the Maximal Likelihood Estimator for intrinsic dimensions from (Levina & Bickel, 2004) in the corrected form due to (MacKay & Zoubin, 2005). The local intrinsic dimension at a point x_i in a dataset $\mathcal{S} \in \mathbb{R}^D$ is approximated by $\hat{d}_k(x_i) = \hat{m}_k(x_i)^{-1}$ where

$$\hat{m}_k(x_i) := \frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(x_i)}{T_j(x_i)} \tag{1}$$

with k being a fixed parameter and $T_j(x_i)$ denoting the distance between x_i and its j-th nearest neighbor. To estimate the dimension of $\mathcal{S}$, the original model from (Levina & Bickel, 2004) uses the average $\frac{1}{\mathcal{S}} \sum_{i=1}^{|\mathcal{S}|} \left(\hat{m}_k(x_i)^{-1}\right)$. The correction in (MacKay & Zoubin, 2005) argues that $\hat{m}_k(x_i)$ is statistically more stable than $\hat{m}_k(x_i)^{-1}$ and proposes to estimate the dimension of $\mathcal{S}$ by

$$\hat{d}_k(\mathcal{S}) := \left(\mathbb{E}_{x_i \in \mathcal{S}} \hat{m}_k(x_i)\right)^{-1}. \tag{2}$$

Because of randomness in sampling, the local dimensions $\hat{m}_k(x_i)$ are not constant across different samples x_i even for an ideal dataset sampled from a probability distribution of smooth density on a manifold. However one can tell how far a dataset $\mathcal{S}$ is from being smoothly distributed on a manifold, by computing the variance of $\hat{m}_k(x_i)$ for $\mathcal{S}$ and comparing it to that of such an ideal dataset. We now introduce a new indicator function that measures this.

Definition 2.1. *The* **manifold indicator** *function of a dataset $\mathcal{S}$ at index k is*

$$I_k(\mathcal{S}) := \frac{(k-1)\mathrm{Var}_{x_i \in \mathcal{S}}(\hat{m}_k(x_i))}{(\mathbb{E}_{x_i \in \mathcal{S}}(\hat{m}_k(x_i)))^2}. \tag{3}$$

Our main conclusion in this section is:

Proposition 2.2. *Under the manifold hypothesis, when $d \ll D$ and $k \ll |\mathcal{S}|$, the distribution of $\hat{m}_k(\mathcal{S})$ satisfies*

$$I_k(\mathcal{S}) \approx 1.$$

The proof of Proposition 2.2 is postponed to Appendix A.2.

For the dataset in §2.1, $I_{20} = 3.55$, $I_{50} = 8.55$, $I_{100} = 17.25$, all of which are much greater than 1. This suggests that the dataset mixes components of different dimensions, consisting with the visualizations in Figures 1-2.

We apply Proposition 2.2 to compare natural image datasets against synthetic datasets smoothly distributed on a connected manifold. The union of manifold hypothesis (Brown et al., 2023) suggested that different connected components of the dataset have different dimensions by observing that the estimated dimensions $\hat{d}(\mathcal{S}_l)$ of different label classes $\mathcal{S}_l$ differ from each other. Our experiments show that even within the same label class, the variance of the local dimension estimate is much higher than what is expected under the manifold hypothesis. This suggests that each connected component has components of different dimensions.

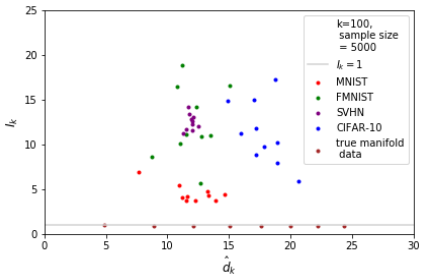

Figure 3. Values of I_{100} for MNIST, FMNIST, SVHN, CIFAR-10

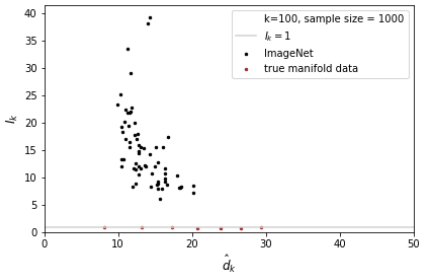

Figure 4. Values of I_{100} for ImageNet

Figures 3-5 compares I_k, at $k = 100$, for individual label classes in MNIST, FMNIST, SVHN, CIFAR-10, CIFAR-100 and ImageNet, compared against the observed I_k for synthetic datasets with smooth densities on manifolds. Each mark in the figures represents an individual label class. For rigorous comparison, all tested classes in each figure, including those of synthetic manifold data, are trimmed to the same sample size N ($N = 5000$ for MNIST, FMNIST, SVHN, CIFAR-10 in Figure 3, $N = 1000$ for ImageNet in 4, and $N = 500$ for CIFAR-100 in 17, due to the smaller size of classes in the last two). For ImageNet, we tested 100 randomly chosen classes.

The smooth manifold data used in the experiments are random points on a randomly generated ellipsoid in $\mathbb{R}^D$. Their sample sizes equal those of other experiments in the same figure. Each figure contains results for several such ellipsoids of different dimensions. The ambient dimension D is the same or close to that of the datasets in the same figure: for Figures 3 and 5 containing MNIST, FMNIST, SVHN, CIFAR-10, CIFAR-100, $D = 3\times32\times32$, for Figures 4 containing ImageNet, each image from ImageNet is rescaled to 224×224 and $D = 3 \times 224 \times 224$.

Results show that synthetic datasets on manifolds do satisfy $I_k \approx 1$ in Proposition 2.2. However, the I_k values for all classes from natural datasets are significantly greater than 1, implying greater variance in $\hat{m}_k$, and hence wider distribution of the local intrinsic dimension $\hat{d}_k(x_i) = \hat{m}_k^{-1}(x_i)$.

Note that large I_k means high variance in local dimensions among different data points. The MNIST dataset is more homogeneous with fewer possible variations in details than

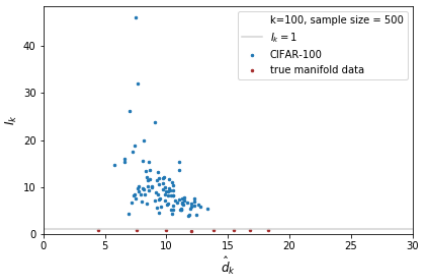

Figure 5. Values of I_{100} for CIFAR-100

other visual objects (e.g. there are fewer ways for people to write the digit "2" than to dress themselves) , thus has relatively less dimension variation among different images compared to SVHN, CIFAR-10, CIFAR-100 or ImageNet. However, since it is significantly bigger than 1, it already reflects a deviation from the manifold hypothesis.

Additional figures recording the same experiments with smaller choices of k are included in Appendix A.2.

3. Interpretation of non-manifold distribution CW complex hypothesis

Evidences from Section 2 suggest that the local intrinsic dimensions may drastically change within the same connected component, contradicting the manifold hypothesis as well as the union of manifolds hypothesis in (Brown et al., 2023). We now outline a few possible reasons behind this phenomenon and suggest an alternative hypothesis.

3.1. Possible reasons of varying dimensions

1. *Invisible features.* An object in an image with s degrees of freedom may be partially or fully blocked by other objects. In this case, only a smaller number $s' < s$ of degrees of freedom can be observed in the image. This results in a local drop of intrinsic dimension by $s - s'$. For example, the rectangle in the geometric dataset in 2.1 can be blocked by the disk and make the image distribution locally degenerate from 2D to 1D. An object may also be partially invisible and cause dimension drop, by being located out of the canvas or being observed by a special angle.

2. *Optional objects.* Objects may not constantly exist in images across the same connected class. For example, portraits may or may not include a mustache or beard of varying size. Objects may have a wide range of decorative details. Whenever such optional features are present, they cause a dimension raise compared to nearby samples without the same features.

3. *Optional global features.* Optional features that add variance to local intrinsic dimensions are not limited to physical objects. For example, pixel values are simultaneously

rescaled by lighting in the scene. The lighting configuration can have many degrees of freedom including the number, colors, brightness, locations, and angles of light source. This configuration interacts with objects in the picture to create shades and highlights. Depending on the interactions, only a subset of the degrees of freedom can be observed in the image. For another example, with strong contrast of lighting, very dark shade is created and objects in the canvas can become invisible. Such interactions change the local intrinsic dimension in data distribution.

4. *Low resolution compared to intrinsic degree of freedom.* An object of s degrees of freedom may be rescaled and placed in different areas of an image. When the area only provides $r < s$ dimensions of resolution, the features describing the object are projected into $\mathbb{R}^r$ and forced to lose their dimension. For example, "a school bus at far" produces only a few pixels colored in yellow, losing all other details.

5. *Concentration near lower dimensional skeletons*. Instead of exact dimension drop, almost drop may also occur when certain degrees of freedom are still visible while making little variation to pixel values. As an example, imagine a picture of sunset where the sun has almost completely sunk below the horizon except its tip. In this case, varying the radius of the sun would change the observed pixel values in the tip but not by a lot. Another example is that changes of objects in very dark shades still affect pixel values but only very slightly. In these situations, the local distribution is still higher dimensional but they are concentrated near a lower dimensional skeleton. This phenomenon can be observed in Figure 1 in the areas where the 1D skeletons are connected to the 2D part, where the thickening takes place gradually.

We validate each of these reasons on a simple synthetic dataset by measuring the indicator function I_k proposed in Section 2.2. The experiment details are in Appendix B.

3.2. CW complex hypothesis

Supported by the empirical evidences and reasons above, we propose the alternative hypothesis that natural image data are supported near unions of "manifolds with skeleton". Such structures are formally defined as CW complexes, an important notion in topology whose name stands for "**C**losure-finite **W**eak topology" (Wikipedia, 2024).

Definition 3.1. *(e.g. (Lundell & Weingram, 2012)) A CW complex X is a Hausdorff topological space, equipped with a disjoint decomposition $X = \bigcup_{j=0}^{d} X^{(j)}$, such that each $X^{(j)}$ is further decomposed into $X^{(j)} = \bigcup_{C\in\mathcal{C}^{(j)}} C$ where:*

(i) Every $C \in C^{(j)}$ is homeomorphic to the j-dimensional open ball B^j via a homeomorphism $\phi_C : B^j \to C$;

(ii) ϕ_C extends to a homeomorphism defined on the closed ball $\overline{B^j}$, and the image of $\phi(\partial B^j)$ is a disjoint union of finitely many lower dimensional cells from $\bigcup_{j'=0}^{j-1} \mathcal{C}^{(j')}$;

(iii)X is equipped with the topology induced by pasting together the topologies on its cells.

The subset $X^{(j)}$ is called the j-dimensional skeleton of X. When the maps ϕ_C are smooth diffeomorphisms, we say X is a smooth CW complex.

As any bounded manifold with smooth boundaries can be decomposed as a union of topological balls, one can use bounded manifolds instead of balls in the definition by adding more lower dimensional cells further dividing existing cells into balls. So we have the following heuristic summary (see illustration in Figure 6):

Summary of Definition 3.1. *A CW complex is the union of a finite family of bounded manifolds (a.k.a. cells), where each cell's boundary consists of finitely many other (lower-dimensional) cells.*

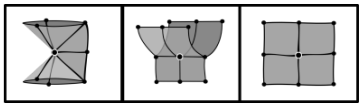

Figure 6. Illustration of CW complices. Source: (Nanda, 2020)

We now formulate the alternative hypothesis.

Hypothesis 3.2 (CW complex hypothesis)**.** *For a generic image dataset $\mathcal{S} \subset \mathbb{R}^D$, its data distribution ν is, up to small perturbation, supported on a union of connected CW-complexes X_i. The restriction $\nu|_{X_i}$ on each connected component X_i is a sum of measures $\sum_{C\in\bigcup_{j=0}^{\dim X_i} \mathcal{C}_i^{(j)}} \nu_C$, where each ν_C is a measure of smooth nonnegative density supported on a cell C of X_i. Note that for a fixed X_i, its cells may have different dimensions.*

4. Consequences of the CW complex hypothesis in deep generative models

4.1. Compatibility with deep neural networks

We now prove Hypothesis 3.2 is highly compatible with the modern deep learning framework, as deep neural networks with ReLu activation automatically reshape smooth probability distributions in parameter spaces into CW complex-supported probability distributions satisfying Hypothesis 3.2. This in particular means that deep generative models (DGMs) are a priori capable of learning distributions of image data under the hypothesis.

Proposition 4.1. *Suppose $F : \mathbb{R}^n \to \mathbb{R}^m$ is a neural network with the ReLu activation function. Then the image set of $F(\mathbb{R}^n)$ is a CW complex. Moreover, if ν_0 is a probability measure of smooth density function on $\mathbb{R}^n$, then the pushforward $F_*\nu_0$ of ν_0 by F satisfies Hypothesis 3.2.*

The proof of Proposition 4.1 is delayed to Appendix C.1. We believe that the conclusion is likely true in generic situations for neural networks with real analytic activation functions such as GELU or the sigmoid function, as the only condition needed is that the common zeros sets of finitely many real analytic functions (which are defined as either components of the neural network or their higher order derivatives) are finite unions of submanifolds. This is an open mathematical question, but we expect it to hold in generic cases.

4.2. Obstruction for diffusion models

In the opposite direction, we investigate potential negative consequences stemming from the necessity to substitute manifold hypothesis with the CW complex hypothesis. Especially, we focus on a topic of practical interest in image generation.

Diffusion models, a class of deep generative models, saw great success recently in generating new images mimicking the distribution of existing data sets. One well-recognized drawback of such models is that they often have difficulty accurately generating image areas with a large amount of details; notably, the numbers and positions of fingers and toes are often incorrectly configured in human portraits, see e.g the article (Chayka, 2023) for a discussion in popular media. Sample images containing such errors can be found in Appendix C.2. We analyze this issue based on the CW complex hypothesis. (Lee et al., 2022; 2023) proved score-based generative models can approximate any data distribution. However, that guarantee relies on the assumption that the neural network used by the model can approximate the training target function arbitrarily well. When the training budget is limited, we hypothesize that the neural network would prioritize certain samples over others based on different intrinsic dimensions. Our interpretation is:

The inconsistency of local intrinsic dimensions make the the deep learning models spend more efforts learning the diffusion gradient for the lower dimensional part of data distribution in regions, than that of the higher dimensional part, near where lower dimensional and higher dimensional cells are connected together. This decreases learning accuracy of the diffusion process along the extra dimensions in the higher dimensional features.

In particular, human hands have remarkably high degrees of freedom in their visual representations. In addition, they often only occupy a small part of the image and many of their intrinsic features may not be visible because of posture. This creates a significant local intrinsic dimension increase in image samples with complicated hands compared to nearby samples containing none of less hand features. Such extra features are often not properly learned, leading to unrealistic partially-learned hand configurations in generated images.

4.2.1. Theoretical interpretation

Our analysis takes as model the restriction of a dataset $\mathcal{S} \subset \mathbb{R}^D$ to a local area U and assumes it consists of a higher dimensional part that is smoothly supported on a d dimensional manifold M, and a lower dimensional part smoothly supported on a "skeleton" manifold K of dimension $k < d < D$. K is glued to M as a cell and may or may not be in the boundary of M. (M is decomposed into CW cells, and K is in the boundary of one or more of them.) For simplicity we shall assume K is in the interior of M. We only care about the learning efficiency within the local area. So for simplicity, ignore curvatures and assume M and K are flat in the neighborhood, and identify M with $\mathbb{R}^d$, and K with $\mathbb{R}^k$, where $\mathbb{R}^k \subset \mathbb{R}^d \subset \mathbb{R}^D$. Let ν be a localized probability measure representing the distribution of data inside U. In view of the CW complex hypothesis, assume ν decomposes as

$$\nu = (1-\gamma)\nu_M + \gamma\nu_K \tag{4}$$

where ν_K (the "spine") is a probability measure of smooth density ρ_K on K and ν_M is a probability measure of smooth density ρ_M on M and the support of ν_K (denoted by $\operatorname{supp}\nu_K$ in the following) is a submanifold of ν_M. As we are only interested in the case where $\operatorname{supp}\nu_K$ and $\operatorname{supp}\nu_M$ are not disjoint from each other, without loss of generality, we will assume and 0 is in both $\operatorname{supp}\nu_K$ and $\operatorname{supp}\nu_M$. After rescaling coordinates if necessary, we may assume that U contains the unit ball $B_{\mathbb{R}^D}(0,1)$ in $\mathbb{R}^D$, and

$$\rho_K|_{B_K(0,1)} > 0, \rho_M|_{B_M(0,1)} > 0. \tag{5}$$

We adopt the new coordinate notation

$$u = (u^{(K)}, u^{(M\ominus K)}, u^{\perp}), \tag{6}$$

where $u^{(K)}, u^{(M\ominus K)}, u^{\perp}$ concatenate respectively the coordinates with indices $\leq k$, in $[k+1, d]$, and $> d$.

Diffusion models, such as the DDPM (Sohl-Dickstein et al., 2015; Ho et al., 2020) and the score-based diffusion model (Song & Ermon, 2019; Song et al., 2021), reconstruct the distribution ν by studying an intermediate random path

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \tag{7}$$

where $x_0 \sim \nu$, $\bar{\alpha}_t \in [0,1]$ is a predetermined schedule scheme and ε is an $\mathcal{N}(0,I)$-distributed random noise. In particular, a neural network is used along the noising process to learn the conditional expectation $\tilde{\boldsymbol{\mu}}(x_t,t) := \mathbb{E}_{x_0\sim\nu}\mathbb{E}(x_{t-1}|x_t,x_0)$. The learned approximation to $\tilde{\boldsymbol{\mu}}$ is then used as a drift term determining the flow direction in a backward stochastic PDE in the denoising process. In practice, $\tilde{\boldsymbol{\mu}}$ was shown to be a linear interpolation between x_0 and x_t, and after reparametrizing, the problem was transformed into the simulating the noise ε by a neural network

$\varepsilon_\theta(x_t, t)$ by minimizing the loss function

$$\mathbb{E}_{x_0\sim\nu,\varepsilon\sim\mathcal{N}(0,I)}\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2. \tag{8}$$

The **intrinsic target noise** $\bar{\varepsilon}(x_t, t)$ is

$$\bar{\varepsilon}(x_t, t) = \mathbb{E}_{x_0\sim\nu,\varepsilon\sim\mathcal{N}(0,I)}\big(\varepsilon\big|x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon\big). \tag{9}$$

By expanding the square in (8), we know that for each given t, minimizing (8) is equivalent to minimizing

$$\mathbb{E}_{x_0\sim\nu,\varepsilon\sim\mathcal{N}(0,I)}\|\bar{\varepsilon}(x_t, t) - \varepsilon_\theta(x_t, t)\|^2. \tag{10}$$

$\bar{\varepsilon}(x_t, t)$ is equivalent to the score function in (Song & Ermon, 2019; Song et al., 2021) by arguments from those papers and (Vincent, 2011).

As pointed out in (Ho et al., 2020) and (Wang & Vastola, 2023), most details in the denoising process are produced when t is small and x_t is close to x_0, while global outlines are generated at the earlier stage when t is larger. Thus the small t stage is largely responsible for determining the local dimension of the data. We claim for small t and $k < j \le d$, the j-th coordinate of the learning target $\bar{\varepsilon}(x_t, t)$ assumes much larger values for x_t near K. For notation simplicity, denote $\delta_t = \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}$ and make a change of variable

$$z_t = \frac{x_t}{\sqrt{\bar{\alpha}_t}} = x_0 + \sqrt{\delta_t}\varepsilon. \tag{11}$$

The following proposition will be proved in Appendix C.3.

Proposition 4.2. *As $\delta_t \to 0$, with the relation (11)*

1. *For the compact domain*

$$\Omega_M = \{z \in M : |z^{(M\ominus K)}| \in [\frac{1}{3}, \frac{2}{3}], |z^{(K)}| \le \frac{1}{2}\}$$

in $M\backslash K$, the $\sqrt{\delta_t}$ neighborhood $\Omega_M^{\sqrt{\delta_t}}$ of Ω_M in the ambient space $\mathbb{R}^d$ satisfies

$$\mathbb{P}_{x_0\sim\nu,\varepsilon\sim\mathcal{N}(0,I)}(z_t \in \Omega_M^{\sqrt{\delta_t}}) \gtrsim 1; \tag{12}$$

$$|\bar{\varepsilon}(x_t, t)^{(M\ominus K)}| \lesssim \sqrt{\delta_t} \text{ if } z_t \in \Omega_M^{\sqrt{\delta_t}}. \tag{13}$$

2. *The cylindrical shell*

$$\Omega_K^{\sqrt{\delta_t}} = \{z \in U : |z^{\perp}| \le \sqrt{\delta_t}, \; |z^{(M\ominus K)}| \in [\frac{\sqrt{\delta_t}}{2}, \sqrt{\delta_t}]\}$$

enclosing K satisfies

$$\mathbb{P}_{x_0\sim\nu,\varepsilon\sim\mathcal{N}(0,I)}(z_t \in \Omega_K^{\sqrt{\delta_t}}) \gtrsim 1; \tag{14}$$

$$|\bar{\varepsilon}(x_t, t)^{(M\ominus K)}| \gtrsim 1 \text{ if } z_t \in \Omega_K^{\sqrt{\delta_t}}. \tag{15}$$

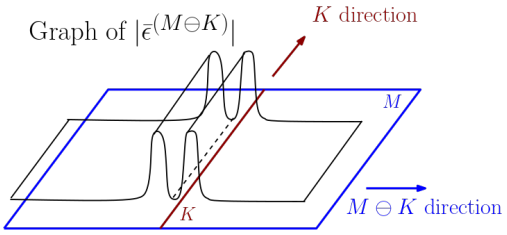

Figure 7. Illustration of Propostion 4.2

The Proposition can be summarized by the illustration in Figure 7. When δ_t is small, the order of magnitude of $\bar{\varepsilon}^{(M\ominus K)}$, the coordinates of the intrinsic learning target vector $\bar{\varepsilon}$ along the extra coordinates $M \ominus K$, is significantly greater near the lower dimensional spine K than near generic regions in the higher dimensional manifold M.

Moreover, these two neighborhoods carry comparable weights in the loss function. This incentivizes the neural network $\varepsilon_\theta(\cdot)$, subject to its training budget, to spend more efforts to learn $\bar{\varepsilon}$ near K and ignore the training along the other tangent directions of M, yielding unsatisfactory results. Because $\bar{\varepsilon}$ is the score vector field used in the denoising process, $\bar{\varepsilon}^{(M\ominus K)}$ is what controls the distribution of generated data along M away from K. This implies that the model may not sufficiently learn the features along extra dimensions of the higher dimensional component M, an example of which would be the details in hands.

It is worth mentioning that in (Wang & Ponce, 2021; Chadebec & Allassonnière, 2022), geometry of deep generative models have been investigated with Riemannian metric tensors explicitly measured on the networks of VAE or GAN. Another related work in that direction is (Wang & Vastola, 2023), which studied how denoising trajectories approach the image manifold and proposes the view that size of the score function is a softmax if the data distribution is a mixture of finitely many Gaussian probabilities. Following that view and making the Gaussians degenerate may provide an alternative proof to Proposition 4.2 in certain cases.

4.2.2. Experiments

Our interpretation is tested by the following experiment steps:

1. Generate a dataset $\mathcal{S} = \mathcal{S}_{\text{high}} \cup \mathcal{S}_{\text{low}}$ consisting of a higher dimensional component $\mathcal{S}_{\text{high}}$ (think of it as ν_M) and a lower dimensional component $\mathcal{S}_{\text{low}}$ (think of it as ν_K), such that the supports of both components are connected, and $\text{supp}\,\mathcal{S}_{\text{low}} \subset \text{supp}\,\mathcal{S}_{\text{high}}$.

2. Train a diffusion model separately on $\mathcal{S}_{\text{high}}$, $\mathcal{S}_{\text{low}}$, $\mathcal{S}$, and generate new datasets $\mathcal{S}'_{\text{high}}$, $\mathcal{S}'_{\text{low}}$, $\mathcal{S}'$ using the model such that $|\mathcal{S}'_{\text{high}}| = |\mathcal{S}_{\text{high}}|$, $|\mathcal{S}'_{\text{low}}| = |\mathcal{S}_{\text{low}}|$, $|\mathcal{S}'| = |\mathcal{S}|$ for easier comparison later.

Figure 8. Samples from $\mathcal{S}_{\text{low}}$ (top row) and $\mathcal{S}_{\text{high}}$ (bottom row)

Table 1. Classification accuracies for generated data.

Classified as:	$\mathcal{S}_{\text{low}}$	$\mathcal{S}_{\text{high}}$
$\mathcal{S}_{\text{low}}$ validation	100%	0%
$\mathcal{S}_{\text{high}}$ validation	0%	100%
$\mathcal{S}'_{\text{low}}$	100%	0%
$\mathcal{S}'_{\text{high}}$	0%	100%
$\mathcal{S}'$	55.84%	44.16%

3. Take the datasets $\mathcal{S}'_{\text{sep\_train}} = \mathcal{S}'_{\text{high}} \cup \mathcal{S}'_{\text{low}}$ (generated by models separately trained on high and low dimensional components) and $\mathcal{S}'$ (generated by model trained with mixed data) and compare them to the original data distribution $\mathcal{S}$ to see which training scheme works better.

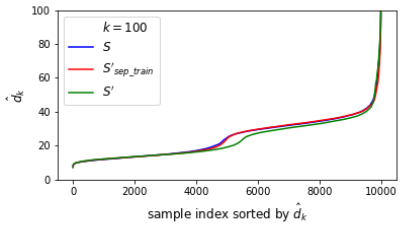

Figure 9. Distribution of $\hat{d}_k$, $k = 100$

Our datasets contain 32×32 images divided into four 16×16 blocks, with every block being a rescaled picture of MNIST digit "2", "4", "6" or "8". For $\mathcal{S}_{\text{low}}$, the pictures "2", "4", "6" are fixed and only the picture "8" is uniformly chosen from MNIST. For $\mathcal{S}_{\text{high}}$, all four pictures "2", "4", "6", "8" are uniformly chosen. Clearly, $\mathcal{S}_{\text{low}}$ has a much lower intrinsic dimension than $\mathcal{S}_{\text{high}}$ as the underlying distribution of the later is the Cartesian product of 4 distributions of individual classes. Moreover, because images in the same digit class are deformations of each other, the data distributions $\mathcal{S}_{\text{low}}$ and $\mathcal{S}_{\text{high}}$ are connected. We use $|\mathcal{S}_{\text{high}}| = |\mathcal{S}_{\text{low}}| = 5000$, and hence $|\mathcal{S}| = 10000$. We use MNIST instead of more complicated datasets in order to generate datasets of the same sample size as the original dataset within training budgets.

We train three DDPM (Ho et al., 2020) models of the same architecture from the open source implementation (Vandegar, 2023) respectively on $\mathcal{S}_{\text{high}}$, $\mathcal{S}_{\text{low}}$, and $\mathcal{S}$. For fair comparison, number of SGD training steps is 50000 for all models, but batch size is doubled from 32 to 64 when training the model on $\mathcal{S}$ so that each image in $\mathcal{S}_{\text{high}}$ is used approximately the same number of times when training on $\mathcal{S}_{\text{high}}$ and $\mathcal{S}$, and similar for $\mathcal{S}_{\text{low}}$.

To compare the generated datasets $\mathcal{S}'_{\text{sep\_train}}$, and $\mathcal{S}'$ against $\mathcal{S}$, we make two tests:

(1) We survey the distribution of local intrinsic dimension estimator (1). For $k = 100$, we sort $\hat{d}_k(x_i)$ for all samples x_i in each of the generated datasets $\mathcal{S}'_{\text{sep\_train}}$, $\mathcal{S}'$ and $\mathcal{S}$ and plot the resulting curves. Figure 9 shows that $\mathcal{S}'_{\text{sep\_train}}$ have similar distribution of $\hat{d}_k$ as $\mathcal{S}$. But the local intrinsic dimension spectra of $\mathcal{S}'$ and $\mathcal{S}$ differ substantially right at the middle of the spectrum where the high dimensional and low dimensional regimes connect to each other and the distribution is twisted towards lower dimensions. Similar results for $k = 50$ and $k = 200$ are included in Appendix C.4.

(2) We train a CNN classifier to distinguish between $\mathcal{S}_{\text{high}}$ and $\mathcal{S}_{\text{low}}$. In test, the classifier is 100% accurate on validation data from the $\mathcal{S}_{\text{high}}$ and $\mathcal{S}_{\text{low}}$, as well as on the separately generated data $\mathcal{S}'_{\text{high}}$ and $\mathcal{S}'_{\text{low}}$. But in $\mathcal{S}'$ generated using mixed data, there is a 5.84% bias favoring $\mathcal{S}_{\text{low}}$. More details, as well as $\hat{d}_k$ statistics of the classier labeled components, can be found in Appendix C.4.

Both tests show that separately training diffusion model on $\mathcal{S}_{\text{high}}$ and $\mathcal{S}_{\text{low}}$ successfully learns the data distributions, but training on the mixed dataset is less accurate. The local intrinsic dimension values $\hat{d}_k$ tend to be lower for $\mathcal{S}'$ compared to $\mathcal{S}$, suggesting diffusion models are more likely to miss higher dimensional features in a portion of samples and has biased generation towards lower-dimensional components, which is consistent with our interpretation.

5. Conclusions, limitations and future directions

We propose the CW complex hypothesis as an alternative to the popular manifold hypothesis and the subsequent union of manifolds hypothesis. We provide several evidences from both theoretical and experimental angles. Our study also exploits the relation with diffusion models and interprets their ineffective learning of higher-dimensional details.

One direction left out in this work is the exploration of possible ways to improve machine learning algorithms in light of the new hypothesis. Though we exposed possible impacts on learning outcomes for image diffusion models, our experiments were on a synthetic dataset constructed out of MNIST that allows a clear distinction between a high dimensional component and a low dimensional one. It would be interesting to extend the study to broader natural datasets

beyond images with varying local intrinsic dimensions and investigate how a better understanding of the variation in intrinsic dimensions can help training models. For instance, we believe efficient clustering of different components of the CW complex would be a future step of potential interest.

Besides consistency of intrinsic dimensions, it is also reasonable to further investigate the other aspect of manifold hypothesis, namely regularity of local geometry.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

The authors are very grateful to the referees for their valuable suggestions and comments.

The first author Y.W. acknowledges the support from her NSF grant.

References

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 6109–6119, 2019.

Arbel, M., Zhou, L., and Gretton, A. Generalized energy based models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Bengio, Y., Courville, A. C., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

Besold, F. and Spokoiny, V. Adaptive manifold clustering. *arXiv:1912.04869*, 2019.

Brand, M. Charting a manifold. In Becker, S., Thrun, S., and Obermayer, K. (eds.), *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, pp. 961–968. MIT Press, 2002.

Brehmer, J. and Cranmer, K. Flows for simultaneous manifold learning and density estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Brown, B. C. A., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. Verifying the union of manifolds hypothesis for image data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Caterini, A. L., Loaiza-Ganem, G., Pleiss, G., and Cunningham, J. P. Rectangular flows for manifold learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 30228–30241, 2021.

Chadebec, C. and Allassonnière, S. A geometric perspective on variational autoencoders. In *Advances in Neural Information Processing Systems*, 2022.

Chayka, K. The uncanny failures of A.I.-generated hands. *The New Yorker*, 2023.

Denti, F., Doimo, D., Laio, A., and Mira, A. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12:20005, 11 2022.

Donoho, D. L. and Grimes, C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences of the United States of America*, 100:5591 – 5596, 2003.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, pp. 983–1049, 2016.

Fodor, I. K. A survey of dimension reduction techniques. In *Technical repor, Lawrence Livermore National Lab., CA (US)*, 2002.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Izquierdo, A. OpenDalle V1.1. https://huggingface.co/dataautogpt3/OpenDalleV1.1, 2023.

Kothari, K., Khorashadizadeh, A., de Hoop, M. V., and Dokmanic, I. Trumpets: Injective flows for inference and inverse problems. In de Campos, C. P., Maathuis, M. H., and Quaeghebeur, E. (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1269–1278. AUAI Press, 2021.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22870–22882. Curran Associates, Inc., 2022.

Lee, H., Lu, J., and Tan, Y. Convergence of score-based generative modeling for general data distributions. In Agrawal, S. and Orabona, F. (eds.), *International Conference on Algorithmic Learning Theory, February 20-23, 2023, Singapore*, volume 201 of *Proceedings of Machine Learning Research*, pp. 946–985. PMLR, 2023.

Levina, E. and Bickel, P. J. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pp. 777–784, 2004.

Lundell, A. T. and Weingram, S. *The topology of CW complexes*. Springer Science & Business Media, 2012.

MacKay, D. J. C. and Zoubin, G. Comments on 'Maximum likelihood estimation of intrinsic dimension' by Levina and Bickel. *The Inference Group Website, Cavendish Laboratory, Cambridge University*, 2005. URL https://www.inference.org.uk/mackay/dimension/.

Marstrand, J. M. Some fundamental geometrical properties of plane sets of fractional dimensions. *Proceedings of The London Mathematical Society*, pp. 257–302, 1954.

Mathieu, E. and Nickel, M. Riemannian continuous normalizing flows. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Nanda, V. Local cohomology and stratification. *Foundations of Computational Mathematics*, 20:195–222, 2020.

Narayanan, H. and Mitter, S. K. Sample complexity of testing the manifold hypothesis. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pp. 1786–1794. Curran Associates, Inc., 2010.

Ozakin, A. and Gray, A. G. Submanifold density estimation. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pp. 1375–1382. Curran Associates, Inc., 2009.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Rezende, D. J., Papamakarios, G., Racaniere, S., Albergo, M., Kanwar, G., Shanahan, P., and Cranmer, K. Normalizing flows on tori and spheres. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *In International Conference on Machine Learning*, pp. 8083—-8092, 2020.

Roweis, S. T. and Saul., L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500): 2323–2326, 2000.

Schölkopf, B., Smola, A. J., and Müller, K. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319, 1998.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. R. and Blei, D. M. (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 2256–2265. JMLR.org, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11895–11907, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Tenenbaum, J. B., Silva, V. D., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Vandegar, M. DDPM implementation code. https://github.com/MaximeVandegar/Papers-in-100-Lines-of-Code/tree/main/Denoising_Diffusion_Probabilistic_Models, 2023.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, 2011.

Wang, B. and Ponce, C. R. The geometry of deep generative image models and its applications. In *International Conference on Learning Representations*, 2021.

Wang, B. and Vastola, J. J. Diffusion models generate images like painters: an analytical theory of outline first, details later. In *https://arxiv.org/abs/2303.02490*, 2023.

Wikipedia. CW complex. https://en.wikipedia.org/wiki/CW_complex, 2024. Accessed: 5 February 2024.

A. Supplemental materials for Section 2

A.1. Additional Experiments for Section 2.1

We document here an experiment that is similar to the one on moving geometric shapes in 2.1. Instead of moving a disk and a rectangle, we construct a synthetic dataset based on MNIST digits. We fix a picture of "2" and another of "4" from MNIST, independently rescale them using scale factors between $\frac{1}{2}$ and 4 randomly chosen using a logarithmic scale. The two resulting pictures are then overlapped and cropped to 16×16 pixels. We again sample 50,000 such randomly generated images and visualize the resulting dataset as in Figures 1 and 2. The sample images and visualizations are in the two figures below:

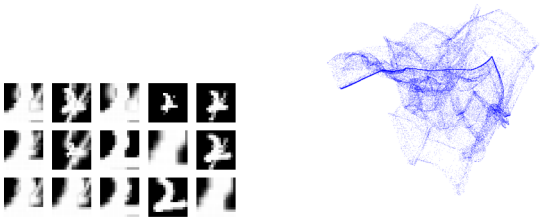

Figure 10. 16x16 random images by overlapping rescaled "2" and "4", and their distributions

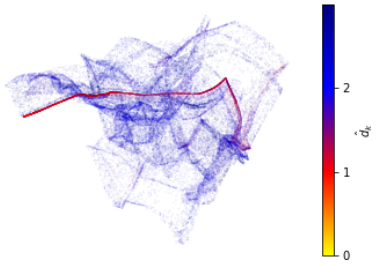

Figure 11. Visualization as in Figure 10, colored based on $\hat{d}_{100}(x_i)$

Like in Figures 1 and 2, the results again show a high concentration near a 1 dimensional curve. Again using $\hat{d}_{100}(x_i) < 1.5$ as a cutoff, the red colored low dimension skeleton contains 16.8% of the dataset. In addition, the I_k value of the datasets are $I_{20} = 2.47$, $I_{50} = 5.68$, $I_{100} = 11.4$, all much greater than 1, which according to Proposition 2.2 confirms the dimension jump observed in the visualization.

A.2. Proof of Proposition 2.2

Assume $\mathcal{S}$ is distributed on a d-dimensional connected smooth manifold $M \subset \mathbb{R}^D$ and is uniformly sampled according to the volume form on M weighted by a smooth density function. It is usually assumed that $d \ll D$ and $k \ll |\mathcal{S}|$. In this regime, the local neighborhood of M containing the k nearest neighbors of a sample x_i can be approximately viewed as a neighborhood inside a linear subspace of dimension d, and the density function is approximately constant on this flat piece. Denoting by $z_j(x_i)$ the j-th nearest neighbor of x_i. then $T_j^d(x_i)$ is proportional to the volume of the ball that is centered at x_i and bounded by the sphere passing through $z_j(x_i)$. It follows that the sequence $T_1^d(x_i), T_2^d(x_i), \cdots$ forms a Poisson process for a fixed x_i when the rest of points are randomly sampled. Based on this fact, it was proved by (Denti et al., 2022) that,

Lemma A.1. *(Denti et al., 2022) Under the manifold hypothesis and when $d \ll D$ and $k \ll |\mathcal{S}|$, the ratio $\frac{T_{l+1}(x_i)}{T_l(x_i)}$ is approximately distributed according to the Pareto law Pareto$(1, ld)$ with scale parameter 1 and exponent ld.*

Moreover, modulo the described approximation by linear spaces, the random variables $\frac{T_l(x_i)}{T_{l-1}(x_i)}$ are jointly independent across different l's.

Proof of Proposition 2.2. Recall that, the distribution Pareto$(1, \alpha)$ is defined by

$$\mathbb{P}_{X\sim\text{Pareto}(1,\alpha)}(X > t) = \begin{cases} t^{-\alpha} & \text{if } t \geq 1; \\ 1 & \text{if } t < 1. \end{cases}$$

Thus, via integration by parts,

$$\begin{aligned} &\mathbb{E}_{X\sim\text{Pareto}(1,\alpha)}(\log X) \\ =&-\int_1^\infty \log t \mathrm{d}\mathbb{P}_{X\sim\text{Pareto}(1,\alpha)}(X > t) \\ =&-(\log t)t^{-\alpha}\big|_{t=1}^\infty + \int_1^\infty (\log t)' t^{-\alpha}\mathrm{d}t \\ =&\int_1^\infty t^{-\alpha-1}\mathrm{d}t = -\alpha^{-1}t^{-\alpha}\big|_{t=1}^\infty = \alpha^{-1}; \end{aligned} \tag{16}$$

and similarly

$$\begin{aligned} &\mathbb{E}_{X\sim\text{Pareto}(1,\alpha)}((\log X)^2) \\ =&-(\log t)^2 t^{-\alpha}\big|_{t=1}^\infty + \int_1^\infty ((\log t)^2)' t^{-\alpha}\mathrm{d}t \\ =&\int_1^\infty 2(\log t)t^{-\alpha-1}\mathrm{d}t \\ =&2\alpha^{-1}\cdot\Big(-\int_1^\infty \log t \mathrm{d}t^{-\alpha}\Big) = 2\alpha^{-2}. \end{aligned}$$

Thus ,

$$\begin{aligned} &\mathrm{Var}_{X\sim\text{Pareto}(1,\alpha)}(\log X) \\ =&\mathbb{E}((\log X)^2) - (\mathbb{E}(\log X))^2 = \alpha^{-2}. \end{aligned} \tag{17}$$

Rewrite (1) as

$$\begin{aligned}\hat{m}_k(x_i) =& \frac{1}{k-1}\sum_{j=1}^{k-1}\sum_{l=j}^{k-1}\log\frac{T_{l+1}(x_i)}{T_l(x_i)} \\ =& \sum_{l=1}^{k-1}\frac{l}{k-1}\log\frac{T_{l+1}(x_i)}{T_l(x_i)}\end{aligned} \tag{18}$$

By using Lemma A.1 and applying (16) and (17) to the independent random variables $\log\frac{T_{l+1}(x_i)}{T_l(x_i)}$, we obtain

$$\mathbb{E}_{x_i\in\mathcal{S}}(\hat{m}_k(x_i)) = \sum_{l=1}^{k-1}\frac{l}{k-1}\cdot\frac{1}{ld} = \frac{1}{d};$$

$$\mathrm{Var}_{x_i\in\mathcal{S}}(\hat{m}_k(x_i)) = \sum_{l=1}^{k-1}(\frac{l}{k-1})^2\cdot\frac{1}{(ld)^2} = \frac{1}{(k-1)d^2},$$

and conclude the proof. □

A.3. Additional Figures for Section 2.2

The Figures below are results for the same experiment as in Figures 3-17 but with smaller choices $k = 20$ and $k = 50$ for the index k, i.e. fewer neighbors are taking into account when computing the local dimension estimator $\hat{d}_k$ and the indicator function I_k. Figures 12-13 correspond to Figure 3; Figures 14-15 correspond to Figure 4, Figures 16-17 correspond to Figure 5. In all tested cases, the indicator value I_k for natural image datasets remain greater than 1, suggesting the manifold hypothesis is not satisfied according to Prop 2.2.

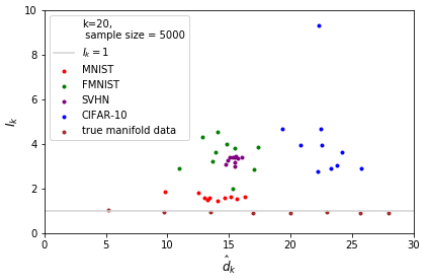

Figure 12. Values of I_{20} for MNIST, FMNIST, SVHN, CIFAR-10

B. Experiments validating possible reasons from Section 3.1

To support the possible reasons listed in §3.1 that lead to varying dimensions, we test the I_k values of several synthetic datasets and compare them to a baseline dataset where none of the reasons are present. All datasets below contain 5000 grayscale images of size 64×64.

"Baseline". Images of a digit "2" and a digit "4", both are fixed choice from MNIST. Both digits are rescaled to

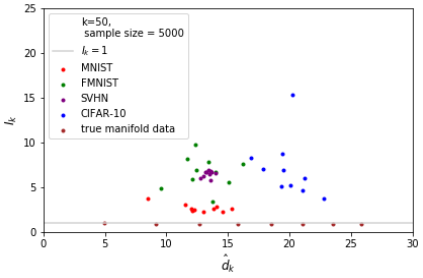

Figure 13. Values of I_{50} for MNIST, FMNIST, SVHN, CIFAR-10

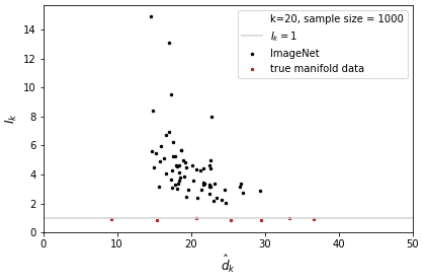

Figure 14. Values of I_{20} for ImageNet

32×32 and slides horizontally, "2" in the top row and "4" in the bottom row. Their horizontal positions are random but never move out of the canvas.

"Block". Similar to "Baseline", but a 32×32 rectangle is attached below the digits "2" and moves accordingly in the bottom row. The square may block the digit "4". This dataset corresponds to Reason 1 in §3.1.

"Out". The digit "2" is allowed to slide out of the canvas through the right side. This dataset corresponds to Reason 1 in §3.1.

"Mesh". A horizontal mesh line separating the top row and the bottom row is added with 50% probability, the line has uniformly random grayscale. This dataset corresponds to Reason 2 in §3.1.

"Lighting". A random lighting scheme is applied to images from the baseline. A random value $a \in [0, 3]$, and the grayscale at a point (x, y) is multiplied by x^a, where the coordinates (x, y) are renormalized to $[0, 1] \times [0, 1]$. This dataset corresponds to Reason 3 in §3.1.

"Shrink". A random shrinking of random scaling factor in $[1/4, 1]$ is applied to images from the baseline. This dataset corresponds to Reason 4 in §3.1.

"Partial Block". Similar to "Block", but a 16×32 rectangle is used instead of a 32×32 rectangle. This new rectangle is not wide enough to block the entire digit "4". This dataset corresponds to Reasons 1 and 5 in §3.1.

Each row in Figure 18 contains sample images from one of the datasets, in the same order as they are listed above.

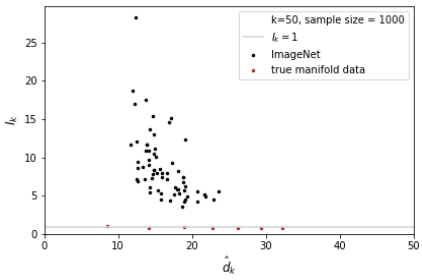

Figure 15. Values of I_{50} for ImageNet

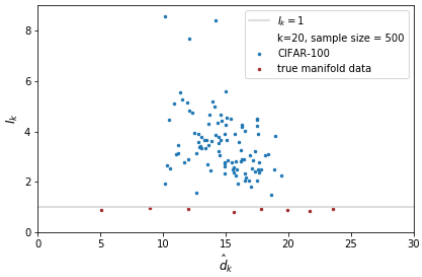

Figure 16. Values of I_{20} for CIFAR-100

Table 2 records the I_k values of the datasets, note that all I_k values increased compared to the baseline. Therefore all the implemented changes increased the variance in local dimensions among data points.

C. Supplementary materials for Section 4

C.1. Proof of Proposition 4.2

Proof. Suppose F has l layers, then it takes the form

$$F = L_l \circ \sigma \circ L_{l-1} \circ \cdots \sigma \circ L_1(x),$$

where $L_i(x) = A_i x + B_i$ are linear functions for $1 \leq i \leq l$, and $\sigma(x) = \mathrm{ReLu}(x)$. ReLu is a piecewise linear function, which is only nonsmooth at 0. Thus F, as a composition of finitely many piecewise linear functions, is a piecewise linear function. More precisely, $\mathbb{R}^n$ is cut into finitely many domains Ω_i by hyperplanes and $F|_{\Omega_i}$ is linear for each i. Its differential DF is a constant matrix on each Ω_i, whose rank is thus a constant on Ω_i but may vary from piece to piece. This implies the image set $F(\Omega_i)$ is a subset of an r-dimensional subspace, where $r = \mathrm{rank} DF|_{\Omega_i}$. Thus the image $F(\mathbb{R}^n)$ is a CW complex whose cells include the sets $F(\Omega_i)$. The boundaries between Ω_i's and their intersecting boundaries, and so on, are also subsets of linear subspaces. The images of them by F are still subsets of linear subspaces and form the other cells of X.

The pushforward measure $F_*\nu_0$ is supported on the $F(\mathbb{R}^n)$ and it preserves the total mass of ν_0. $F_*\nu_0$ restricted to $F(\Omega_i)$ is smooth with respect to the volume measure along the linear subspace $F(\Omega_i)$. The images of the boundaries between Ω_i's and their recursively defined lower dimensional boundaries do not support components of $F_*\nu_0$. □

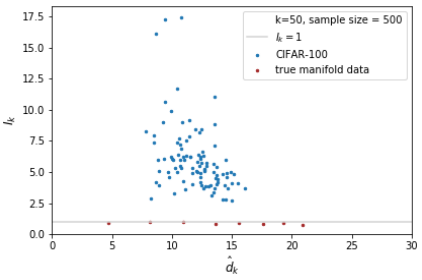

Figure 17. Values of I_{50} for CIFAR-100

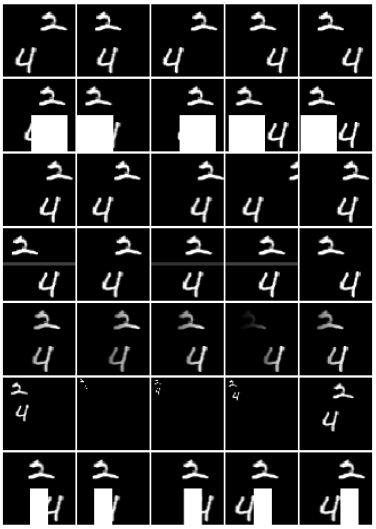

Figure 18. Sample images representing different reasons

C.2. Examples of AI-generated images with wrong details in fingers

Figure 19 contains examples of images with hands were generated by the open-source text-to-image diffusion model OpenDalle V1.1 (Izquierdo, 2023). Despite being highly realistic otherwise, they contain errors in various details of the fingers: number (a, b), texture (c) and shape (d).

C.3. Proof of Proposition 4.2

Denote by $\rho_0(z) = (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}|z|^2}$ the density function of the standard normal distribution $\varepsilon \sim \mathcal{N}(0, I)$ on $\mathbb{R}^D$. Then the density function of $\sqrt{\delta_t}\varepsilon$ is

$$(\sqrt{\delta_t})^{-D} \rho_0(\frac{z}{\sqrt{\delta_t}}). \tag{19}$$

Retain the notations from notations from Proposition 4.2 and relation (11) for all the discussions below. It will always be assumed that $\delta_t \approx 0$. Recall that we identify M with $\mathbb{R}^d$ and K with $\mathbb{R}^k$.

Table 2. I_k values of datasets

Dataset	I_{10}	I_{20}	I_{50}
Baseline	1.04	1.07	1.26
Block	2.33	4.57	21.2
Out	1.94	2.97	5.59
Mesh	1.50	1.67	2.34
Lighting	1.16	1.46	3.04
Shrink	1.57	2.11	3.44
Partial Block	1.12	1.54	5.26

(a) (b) (c) (d)

Figure 19. AI-generated images with mistakes in fingers. Source: OpenDalle V1.1 (Izquierdo, 2023)

Lemma C.1. *For all* $z \in \Omega_M^{\sqrt{\delta_t}}$,

$$\mathbb{E}_{x_0\sim\nu_M}(\sqrt{\delta_t})^{-D}\rho_0(\frac{z-x_0}{\sqrt{\delta_t}}) \gtrsim (\sqrt{\delta_t})^{-(D-d)}, \quad (20)$$

$$\begin{aligned}&\mathbb{E}_{x_0\sim\nu_K}(\sqrt{\delta_t})^{-D}\rho_0(\frac{z-x_0}{\sqrt{\delta_t}})\\ \lesssim& e^{-O(\delta_t^{-1})}(\sqrt{\delta_t})^{-(D-d)},\end{aligned} \quad (21)$$

and

$$\begin{aligned}&\mathbb{E}_{x_0\sim\nu_K}(\sqrt{\delta_t})^{-D}\rho_0(\frac{z-x_0}{\sqrt{\delta_t}})\Big|\frac{z-x_0}{\sqrt{\delta_t}}\Big|\\ \lesssim& e^{-O(\delta_t^{-1})}(\sqrt{\delta_t})^{-(D-d)}.\end{aligned} \quad (22)$$

Proof. By the assumption (5), $\nu_M(\Omega_M) \gtrsim 1$ and ν_M has a uniformly positive density on Ω_M. As $z \in \Omega_M^{\sqrt{\delta_t}}$, the set

$$B(z, 2\sqrt{\delta_t}) \cap M = \{x_0 \in M : |x_0 - z| \le 2\sqrt{\delta_t}\} \quad (23)$$

has at least radius $\gtrsim \sqrt{\delta_t}$ and is located within the unit ball $B_M(0,1)$ in M. Because of assumption (5), the ν_M measure of this set is $\gtrsim (\sqrt{\delta_t})^d$. Furthermore, the value of $\rho_0(\frac{z-x_0}{\sqrt{\delta_t}})$ is $\gtrsim 1$ for all $x_0 \in B(x_0, 2\sqrt{\delta_t}) \cap M$. The inequality (20) then follows by only counting the contribution of $x_0 \in B(x_0, 2\sqrt{\delta_t}) \cap M$.

For (21), notice that if $x_0 \in \operatorname{supp}\nu_K \subset K$ and $z \in \Omega_M^{\delta_t}$, then $|z - x_0| \ge \frac{1}{3} - 2\sqrt{\delta_t} \ge \frac{1}{6}$. Hence $\frac{z-x_0}{\sqrt{\delta_t}} \gtrsim \frac{1}{\sqrt{\delta_t}}$ and

$$\rho_0(\frac{z-x_0}{\sqrt{\delta_t}}) \lesssim e^{-O(\delta_t^{-1})}.$$

After integrating and adjusting the implicit coefficient in $O(\delta_t^{-1})$, this implies (21). Again by adjusting the implicit coefficient, we also have

$$\rho_0(\frac{z-x_0}{\sqrt{\delta_t}})\frac{1}{\sqrt{\delta_t}} \lesssim e^{-O(\delta_t^{-1})},$$

which leads to (22). □

Corollary C.2. *The set* $\Omega_M^{\sqrt{\delta_t}}$ *satisfies*

$$\mathbb{P}_{x_0\sim\nu_M,\varepsilon\sim\mathcal{N}(0,I)}(z_t \in \Omega_M^{\sqrt{\delta_t}}) \gtrsim 1,$$

and

$$\mathbb{P}_{x_0\sim\nu_K,\varepsilon\sim\mathcal{N}(0,I)}(z_t \in \Omega_M^{\sqrt{\delta_t}}) \lesssim e^{-O(\delta_t^{-1})}.$$

Proof. Notice that the $\operatorname{vol}_M(\Omega_M) \asymp 1$ and hence $\operatorname{vol}(\Omega_M^{\sqrt{\delta_t}}) \asymp (\sqrt{\delta_t})^{(D-d)}$. The rest of proof is straightforward, by integrating Lemma C.1. □

Lemma C.3. *For all* $z \in U$

$$\left|\mathbb{E}_{x_0\sim\nu_M,\varepsilon\sim\mathcal{N}(0,I)}(\varepsilon^{(M)}\big|z_t = z)\right| \lesssim \sqrt{\delta_t}$$

Proof. Following the coordinate system (6), we write $z = (z^{(K)}, z^{(M\ominus K)}, z^\perp) = (z^{(M)}, z^\perp)$. Similarly, we decompose the noise ε as $(\varepsilon^{(K)}, \varepsilon^{(M\ominus K)}, \varepsilon^\perp) = (\varepsilon^{(M)}, \varepsilon^\perp)$, where $z^{(M)}$ and $\varepsilon^{(M)}$ are the projections to $M = \mathbb{R}^d$. Then

$$\begin{aligned}&\mathbb{E}_{x_0\sim\nu_M,\varepsilon\sim\mathcal{N}(0,I)}\big(\varepsilon^{(M)}\big|z_t = z\big)\\ =&\mathbb{E}_{x_0\sim\nu_M,\varepsilon\sim\mathcal{N}(0,I)}\\ &\big(\varepsilon^{(M)}\big|\sqrt{\delta_t}\varepsilon^\perp = z^\perp, x_0 + \sqrt{\delta_t}\varepsilon^{(M)} = z^{(M)}\big)\\ =&\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\rho_M(z^{(M)} - \sqrt{\delta_t}\varepsilon^{(M)})\varepsilon^{(M)}.\end{aligned}$$

Here we used the fact that $\varepsilon^\perp$ and $\varepsilon^{(M)}$ are independent noises. By Taylor expansion, the expression above further equals

$$\begin{aligned}&\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\rho_M(z^{(M)} - \sqrt{\delta_t}\varepsilon^{(M)})\varepsilon^{(M)}\\ =&\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\Big[\rho_M(z^{(M)})\varepsilon^{(M)} +\\ &+\nabla_M\rho_M(z^{(M)})(\sqrt{\delta_t}\varepsilon^{(M)})^\top\varepsilon^{(M)}\\ &+O(\delta_t)\Big],\end{aligned}$$

where ∇_M denotes the gradient operator along M.

Note that as $\varepsilon \sim \mathcal{N}(0, I)$ on $\mathbb{R}^D$, $\mathbb{E}\varepsilon^{(M)} = (\mathbb{E}\varepsilon)^{(M)} = 0$. It follows that $\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}(\rho_M(z^{(M)})\varepsilon^{(M)}) = \rho_M(z^{(M)})\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}(\varepsilon^{(M)}) = 0$, and the expression at hand is of order $O(\sqrt{\delta_t})$. This proves the lemma. □

Lemma C.4. *For all $z \in \Omega_K^{\sqrt{\delta_t}}$,*

$$\mathbb{E}_{x_0\sim\nu_M}(\sqrt{\delta_t})^{-D}\rho_0(\frac{z-x_0}{\sqrt{\delta_t}}) \lesssim (\sqrt{\delta_t})^{-(D-d)}, \quad (24)$$

$$\mathbb{E}_{x_0\sim\nu_K}(\sqrt{\delta_t})^{-D}\rho_0(\frac{z-x_0}{\sqrt{\delta_t}}) \gtrsim (\sqrt{\delta_t})^{-(D-k)}. \quad (25)$$

Proof. The proof of (25) is similar to that of (20) in Lemma C.1 and is omitted. That of (24) is more delicate and is presented below.

Define $B(z, 2\sqrt{\delta_t}) \cap M$ as in (23). The volume in M, and hence ν_M measure, of $B(z, 2\sqrt{\delta}) \cap M$, is $O((\sqrt{\delta_t})^d)$. Thus the contribution of $x_0 \in B(z, 2\sqrt{\delta_t}) \cap M$ to (24) is $\lesssim (\sqrt{\delta_t})^d \cdot (\sqrt{\delta_t})^{-D} = (\sqrt{\delta_t})^{-(D-d)}$ because $|\rho_0| \le 1$. It remains to estimate the contribution of the complement set $\{x_0 \in M : |z - x_0| \ge 2\sqrt{\delta}\}$. For such x_0, $|z - x_0| \ge \frac{1}{2}|z^{(M)} - x_0|$, where $z^{(M)}$ is the projection of z to M in the coordinate system (6). Therefore, $\rho_0(\frac{z-x_0}{\sqrt{\delta_t}}) \le \rho_0(\frac{z^{(M)}-x_0}{2\sqrt{\delta_t}})$. Hence the contribution of this part is bounded by

$$\begin{aligned}
&\mathbb{E}_{x_0\sim\nu_M}(\sqrt{\delta_t})^{-D}\rho_0(\frac{z^{(M)}-x_0}{2\sqrt{\delta_t}}) \\
\lesssim&(\sqrt{\delta_t})^{-(D-d)}\mathbb{E}_{x_0\sim\nu_M}(\sqrt{\delta_t})^{-d}\rho_0(\frac{z^{(M)}-x_0}{2\sqrt{\delta_t}}) \\
\lesssim&(\sqrt{\delta_t})^{-(D-d)}\int_M(\sqrt{\delta_t})^{-d}\rho_0(\frac{z^{(M)}-x_0}{2\sqrt{\delta_t}})\mathrm{d}x_0 \\
\lesssim&(\sqrt{\delta_t})^{-(D-d)}.
\end{aligned}$$

Here we used the fact that ν_M has smooth density on M and is hence bounded by a multiple of the volume form on M, and the fact that $\int_M(\sqrt{\delta_t})^{-d}\rho_0(\frac{z^{(M)}-x_0}{2\sqrt{\delta_t}})\mathrm{d}x_0$ is an absolute constant (this is because an easy computation shows that $(2\pi)^{\frac{D-d}{2}}(2\sqrt{\delta_t})^{-d}\rho_0(\frac{z^{(M)}-x_0}{2\sqrt{\delta_t}})$ is the density function of the normal distribution on $M = \mathbb{R}^d$, with standard deviation $2\sqrt{\delta_t}$ centered at $z^{(M)}$). The proof is completed. □

Corollary C.5. *The set $\Omega_K^{\sqrt{\delta_t}}$ satisfies*

$$\mathbb{P}_{x_0\sim\nu_K,\varepsilon\sim\mathcal{N}(0,I)}(z_t \in \Omega_K^{\sqrt{\delta_t}}) \gtrsim 1.$$

Proof. Like in Corollary C.2, this is established by integrating the previous lemma using $\mathrm{vol}(\Omega_K^{\sqrt{\delta_t}}) \asymp (\sqrt{\delta_t})^{(D-k)}$. □

Lemma C.6. *For all $z \in \Omega_K^{\sqrt{\delta_t}}$,*

$$\left|\mathbb{E}_{x_0\sim\nu_K,\varepsilon\sim\mathcal{N}(0,I)}(\varepsilon^{(M\ominus K)}\big|z_t = z)\right| \asymp 1$$

Proof. For generic $x_0 \sim \nu_K$ and $\varepsilon \sim \mathcal{N}(0, I)$, if $z_t = x_0 + \sqrt{\delta_t}\varepsilon$ is equal to z, then $z^{(M\ominus K)} = \sqrt{\delta_t}\varepsilon^{(M\ominus K)}$ because $x_0 \in \mathrm{supp}\, K = K$ projects to 0 along the $(M \ominus K)$ coordinates. Hence the left hand side is equal to $\frac{z^{(M\ominus K)}}{\sqrt{\delta_t}}$, which is in $[\frac{1}{3}, \frac{2}{3}]$ by the definition of $\Omega_K^{\sqrt{\delta_t}}$. □

Proof of Proposition 4.2. As $\nu = (1-\gamma)\nu_M + \gamma\nu_K$, by Bayes' rule,

$$\begin{aligned}
&\bar{\varepsilon}(x_t, t)^{(M\ominus K)} \\
=&\mathbb{E}_{x_0\sim\nu,\varepsilon\sim\mathcal{N}(0,I)}\big(\varepsilon^{(M\ominus K)}\big|z_t = x_0 + \sqrt{\delta_t}\varepsilon\big) \\
=&\Big[(1-\gamma)\mathbb{E}_{x_0\sim\nu_M}(\sqrt{\delta_t})^{-D}\rho_0(\frac{z_t-x_0}{\sqrt{\delta_t}}) \\
&\cdot\mathbb{E}_{x_0\sim\nu_M,\varepsilon\sim\mathcal{N}(0,I)}\big(\varepsilon^{(M\ominus K)}\big|z_t = x_0 + \sqrt{\delta_t}\varepsilon\big) \\
&+\gamma\mathbb{E}_{x_0\sim\nu_K}(\sqrt{\delta_t})^{-D}\rho_0(\frac{z_t-x_0}{\sqrt{\delta_t}}) \\
&\cdot\mathbb{E}_{x_0\sim\nu_K,\varepsilon\sim\mathcal{N}(0,I)}\big(\varepsilon^{(M\ominus K)}\big|z_t = x_0 + \sqrt{\delta_t}\varepsilon\big)\Big] \\
&\Big/\Big[(1-\gamma)\mathbb{E}_{x_0\sim\nu_M}(\sqrt{\delta_t})^{-D}\rho_0(\frac{z_t-x_0}{\sqrt{\delta_t}}) \\
&+\gamma\mathbb{E}_{x_0\sim\nu_K}(\sqrt{\delta_t})^{-D}\rho_0(\frac{z_t-x_0}{\sqrt{\delta_t}})\Big]
\end{aligned} \quad (26)$$

1. The case of $z_t \in \Omega_M^{\sqrt{\delta_t}}$. The inequality (12) is given by Corollary (C.2), and it remains to prove (13). By the inequalities in Lemmas C.1 to C.6, the numerator in (26) has upper bound

$$\begin{aligned}
|\text{numerator}| \lesssim&(1-\gamma)(\sqrt{\delta_t})^{-(D-d)}\sqrt{\delta_t} \\
&+\gamma e^{-O(\delta_t^{-1})}(\sqrt{\delta_t})^{-(D-d)} \\
\lesssim&(\sqrt{\delta_t})^{-(D-d)}\sqrt{\delta_t}.
\end{aligned}$$

And denominator has lower bound

$$\begin{aligned}
|\text{denominator}| \gtrsim&(1-\gamma)(\sqrt{\delta_t})^{-(D-d)} \\
&-\gamma O\big(e^{-O(\delta_t^{-1})}(\sqrt{\delta_t})^{-(D-d)}\big) \\
\gtrsim&(\sqrt{\delta_t})^{-(D-d)}
\end{aligned}$$

We obtain (13) by taking quotient.

2. The case of $z_t \in \Omega_K^{\sqrt{\delta_t}}$. Again, Corollary (C.5) gives (14) and we now prove (15). This time, by Lemma C.4, the second terms in the denominator is dominating as $k < d$, hence one can ignore the first term. In other words, it suffices to prove

$$\left| \frac{1-\gamma}{\gamma} \frac{\mathbb{E}_{x_0\sim\nu_M}(\sqrt{\delta_t})^{-D}\rho_0(\frac{z_t-x_0}{\sqrt{\delta_t}})}{\mathbb{E}_{x_0\sim\nu_K}(\sqrt{\delta_t})^{-D}\rho_0(\frac{z_t-x_0}{\sqrt{\delta_t}})} \cdot \mathbb{E}_{x_0\sim\nu_M,\varepsilon\sim\mathcal{N}(0,I)}\left(\varepsilon^{(M\ominus K)}\middle| z_t = x_0+\sqrt{\delta_t}\varepsilon\right) + \mathbb{E}_{x_0\sim\nu_K,\varepsilon\sim\mathcal{N}(0,I)}\left(\varepsilon^{(M\ominus K)}\middle| z_t = x_0+\sqrt{\delta_t}\varepsilon\right) \right| \gtrsim 1. \tag{27}$$

Which is true because the second term's norm is $\gtrsim 1$ by Lemma C.6 and that the first term's norm has an upper bound of order

$$\lesssim \frac{(\sqrt{\delta_t})^{-(D-d)}}{(\sqrt{\delta_t})^{-(D-k)}} \cdot \sqrt{\delta_t} \ll 1$$

by Lemma C.3 and C.4. □

C.4. Additional information on experiments from §4.2.2

The next two figures complements Figure 9 with two other choices of k.

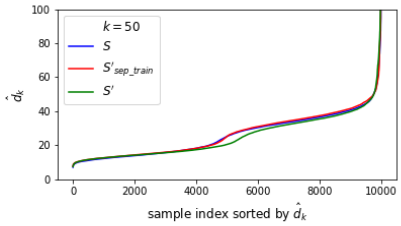

Figure 20. Distribution of $\hat{d}_k$, $k = 50$

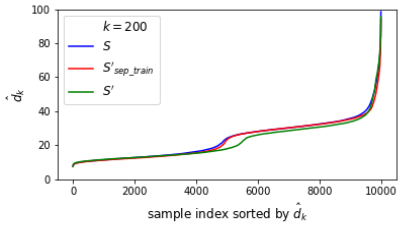

Figure 21. Distribution of $\hat{d}_k$, $k = 200$

Implementation of the classifier in Table 1. The classifier is a convolutional neural network with 2 convolution layers and 2 linear layers and has 24,502 parameters. $3,000$ samples from each of $\mathcal{S}_{\text{low}}$ and $\mathcal{S}_{\text{high}}$ are used as training data respectively for labels "low" and "high", and the rest are used as validation data.

Dimension statistics of generated data after classification. The classifier labeled 5584 images from $\mathcal{S}'_{\text{sep\_train}}$ as $\mathcal{S}_{\text{low}}$ and 4416 as $\mathcal{S}_{\text{high}}$, we denote these two sets respectively as $\mathcal{S}'_{\text{sep\_train,low}}$, and $\mathcal{S}'_{\text{sep\_train,high}}$. It turns out the distributions of local intrinsic dimensions of these two classifier-labeled sets are similar to those of $\mathcal{S}_{\text{low}}$ and $\mathcal{S}_{\text{high}}$, respectively.

Because $|\mathcal{S}'_{\text{sep\_train,low}}| = 5584$ and $|\mathcal{S}'_{\text{sep\_train,high}}| = 4416$ but $|\mathcal{S}_{\text{low}}| = |\mathcal{S}_{\text{high}}| = |\mathcal{S}'_{\text{low}}| = |\mathcal{S}'_{\text{high}}| = 5000$, for equal comparison we randomly truncate the datasets so that they all have size 4000, while keeping their names without further notice. Local intrinsic dimension estimates $\hat{d}_k(x_i)$ are measured for each of these datasets. These estimates count only neighbors from the same dataset, instead of the larger mixed datasets $\mathcal{S}$, $\mathcal{S}'$ or $\mathcal{S}'_{\text{sep\_train}}$. Plots in Figures 22-27 of sorted values of $\hat{d}_k$ show that $\mathcal{S}'_{\text{sep\_train,low}}$, $\mathcal{S}'_{\text{low}}$ and $\mathcal{S}_{\text{low}}$ share similar intrinsic dimension statistics, and so do $\mathcal{S}'_{\text{sep\_train,high}}$, $\mathcal{S}'_{\text{high}}$ and $\mathcal{S}_{\text{high}}$.

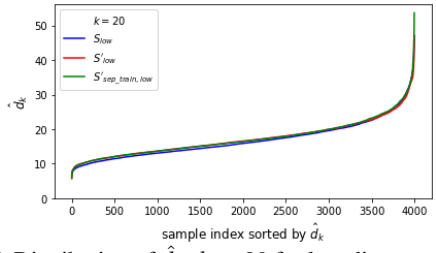

Figure 22. Distribution of $\hat{d}_k$, $k = 20$ for low dim components

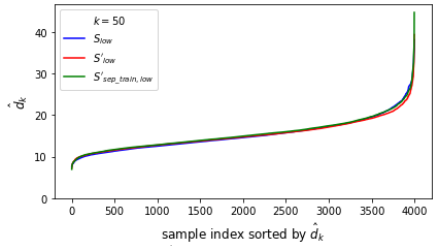

Figure 23. Distribution of $\hat{d}_k$, $k = 50$ for low dim components

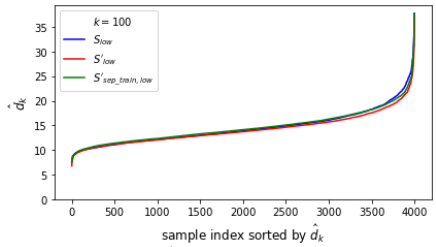

Figure 24. Distribution of $\hat{d}_k$, $k = 100$ for low dim components

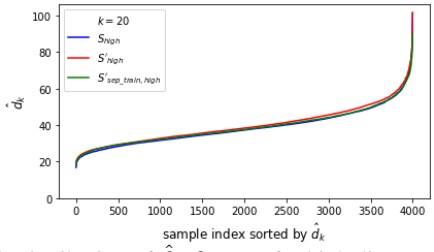

Figure 25. Distribution of $\hat{d}_k$, $k = 20$ for high dim components

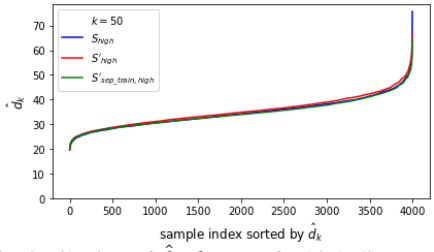

Figure 26. Distribution of $\hat{d}_k$, $k = 50$ for high dim components

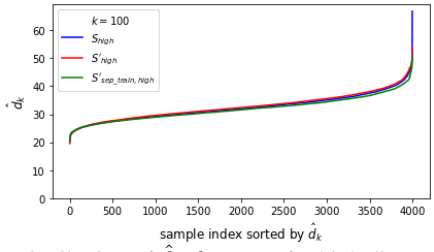

Figure 27. Distribution of $\hat{d}_k$, $k = 100$ for high dim components