Diffusion Models Encode the Intrinsic Dimension of the Data Manifolds

[1]Jan Stanczuk*, [1]Georgios Batzolis*,
[2]Teo Deveney, [1]Carola-Bibiane Schönlieb

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge
[2]Department of Mathematical Sciences, University of Bath
*Equal contribution

Set-up

Problem set-up:

- There is a *data manifold* $\mathcal{M}$ of *intrinsic dimension* k embedded in an ambient euclidean space $\mathbb{R}^d$. Where $d \gg k$.
- There is a probability distribution p, which is highly concentrated around $\mathcal{M}$.
- We are given a finite sample of data $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^{\mathrm{d}}$ generated from p.

Diffusion models are generative models designed to learn p, but they don't explicitly find k. We show how one could try to extract k from an already trained diffusion model.

Diffusion models and the score function

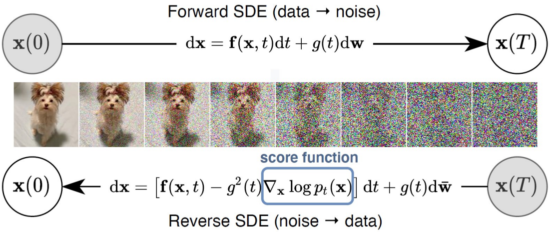
We consider an Ito's diffusion:

$$dx = f(x, t)dt + g(t)dW$$

We can obtain time reversal of the SDE:

$$dx = [f(x, t) - g(t)^2 \nabla_x \ln p_t(x)]dt + g(t)dW$$

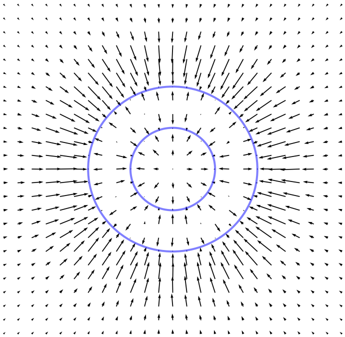
We use a neural network $s_\theta(x, t)$ to approximate the score function $\nabla_x \ln p_t(x)$.



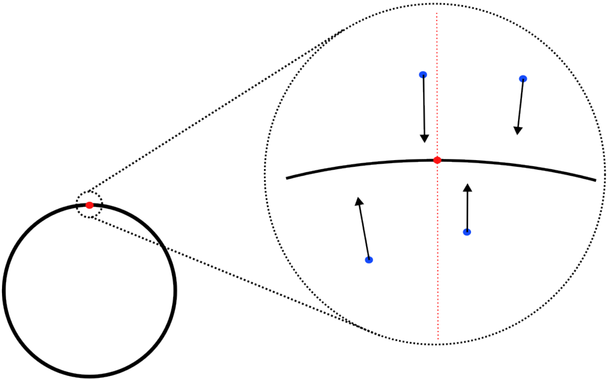
Y.Song et.al.

Score field is perpendicular to the manifold

$$s_\theta(x, \varepsilon) \approx \nabla_x \ln p_\varepsilon(x)$$



Locally score vectors lie in the normal space



Intrinsic dimension estimation method

Estimate the intrinsic dimension at $\mathbf{x}_0$

Input: s_θ - trained diffusion model (score)
t_0 - sampling time
K - number of score vectors.

Sample $\mathbf{x}_0 \sim p_0(\mathbf{x})$ from the data set
$d \leftarrow \dim(\mathbf{x}_0)$
$S \leftarrow$ empty matrix
for $i = 1, ..., K$ **do**
    Sample $\mathbf{x}_{t_0}^{(i)} \sim \mathcal{N}(\mathbf{x}_{t_0} | \mathbf{x}_0, \sigma_{t_0}^2 \mathbf{I})$
    Append $s_\theta(\mathbf{x}_{t_0}^{(i)}, t_0)$ as a new column to S
$(s_i)_{i=1}^d, (\mathbf{v}_i)_{i=1}^d, (\mathbf{w}_i)_{i=1}^d \leftarrow \mathrm{SVD}(S)$
$\hat{k}(\mathbf{x}_0) \leftarrow d -_{i=1,..,d-1} (s_i - s_{i+1})$
Output: $\hat{k}(\mathbf{x}_0)$

Theory

Theorem

Let the support of the data distribution P_0 be contained within a compact embedded sub-manifold $\mathcal{M} \subseteq \mathbb{R}^d$. Denote by P_t the distribution of samples from P_0 diffused over a time duration t.
For any $\mathbf{x} \in \mathbb{R}^d$ sufficiently close to $\mathcal{M}$ with its orthogonal projection on $\mathcal{M}$ denoted as $\pi(\mathbf{x})$, if $\mathbf{n}$ is a unit vector pointing from $\mathbf{x}$ towards $\pi(\mathbf{x})$, then under mild conditions, for any unit vector $\boldsymbol{\nu}$ orthogonal to $\mathbf{n}$, the following holds:

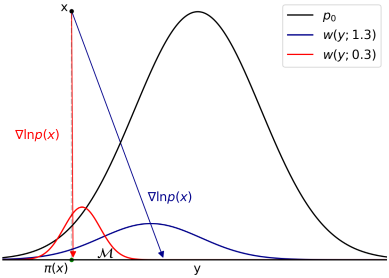
$$\frac{\boldsymbol{\nu}^T \nabla_{\boldsymbol{x}} \ln p_t(\boldsymbol{x})}{\boldsymbol{n}^T \nabla_{\boldsymbol{x}} \ln p_t(\boldsymbol{x})} \to 0 \quad \text{as } t \to 0.$$

Illustrative Simple Case - line embedded in $\mathbb{R}^2$

The score at point $\mathbf{x}$ is given by

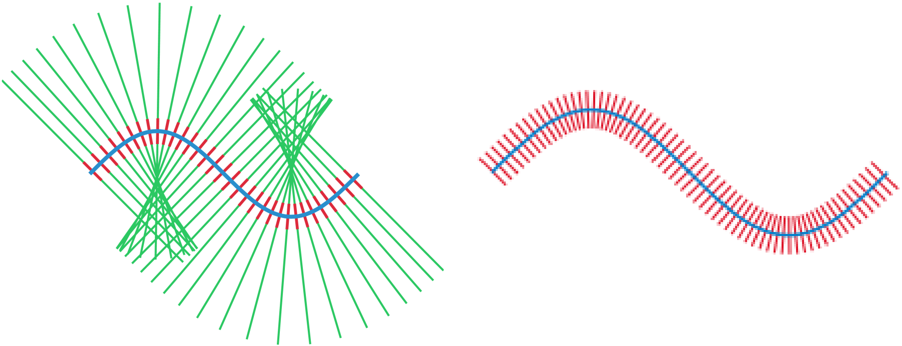
$$\nabla_{\mathbf{x}} \ln p_t(\mathbf{x}) = \frac{1}{\sigma_t^2 p_t(\mathbf{x})} \int_{\mathcal{M}} (\mathbf{y} - \mathbf{x}) \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma_t^2 \mathbf{I}) p_0(\mathbf{y}) d\mathbf{y}$$
$$= \frac{1}{\sigma_t^2 p_t(\mathbf{x})} \int_{\mathcal{M}} (\mathbf{y} - \mathbf{x}) w_x(y; \sigma_t) d\mathbf{y}.$$

The score is the weighted average of vectors pointing from $\mathbf{x}$ to $\mathbf{y}$ with weights given by $w_{\mathbf{x}}(\mathbf{y}; \sigma_t)$. As σ_t decreases, $w_{\mathbf{x}}(\mathbf{y}; \sigma_t)$ concentrates around $\pi(\mathbf{x})$. Therefore, the score direction aligns with $\pi(\mathbf{x}) - \mathbf{x}$ as $\sigma_t \to 0$



Sketch of proof

The tubular neighborhood is a band around the manifold that contains all points with a unique projection on the manifold (shown below in red).



Every compact embedded submanifold of $\mathbb{R}^d$ has a tubular neighborhood.

Sketch of proof (continued)

Let $f_{\mathbf{x}}(\mathbf{y}) : \mathcal{M} \to \mathbb{R}$ denote the squared distance function from $\mathbf{x}$ given by $f_{\mathbf{x}}(\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$.
Using Morse theory, we establish that if $\mathbf{x}$ is in a tubular neighbourhood:

1. $\pi(\mathbf{x})$ is a non-degenerate critical point of $f_{\mathbf{x}}$ of index zero.
2. There exists a connected open neighbourhood E of $\pi(\mathbf{x})$ such that

$$\forall_{\mathbf{y} \in E} \forall_{\tilde{\mathbf{y}} \in \mathcal{M} \setminus E} f_{\mathbf{x}}(\mathbf{y}) < f_{\mathbf{x}}(\tilde{\mathbf{y}}). \tag{1}$$

Lemma

There exists a connected open neighbourhood E of $\pi(\mathbf{x})$ such that,

$$\frac{\int_{\mathcal{M} \setminus E} \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma_t^2 \mathbf{I}) d\mathbf{y}}{\int_E \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma_t^2 \mathbf{I}) d\mathbf{y}} \to 0 \textit{ as } t \to 0. \tag{2}$$

Sketch of proof (continued)

- Consider:

$$\frac{\int_{\mathcal{M}\setminus E} \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma_t^2\mathbf{I})d\mathbf{y}}{\int_E \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma_t^2\mathbf{I})d\mathbf{y}}$$

- By the Mean Value Theorem:

$$\int_E \exp\{-f_{\mathbf{x}}(\mathbf{y})/2\sigma_t^2\}d\mathbf{y} = \mathrm{Vol}(E)\exp\{-f_{\mathbf{x}}(\mathbf{y}^*)/2\sigma_t^2\}$$

$$\int_{\mathcal{M}\setminus E} \exp\{-f_{\mathbf{x}}(\mathbf{y})/2\sigma_t^2\}d\mathbf{y} = \mathrm{Vol}(\mathcal{M}\setminus E)\exp\{-f_{\mathbf{x}}(\tilde{\mathbf{y}}^*)/2\sigma_t^2\}$$

- Result:

$$\frac{\int_{\mathcal{M}\setminus E} \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma_t^2\mathbf{I})d\mathbf{y}}{\int_E \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma_t^2\mathbf{I})d\mathbf{y}} = \frac{\mathrm{Vol}(\mathcal{M}\setminus E)}{\mathrm{Vol}(E)}\exp\left\{-\frac{f_{\mathbf{x}}(\tilde{\mathbf{y}}^*) - f_{\mathbf{x}}(\mathbf{y}^*)}{2\sigma_t^2}\right\}$$

Main theoretical result

Corollary

The ratio of the projection of the score $\nabla_{\boldsymbol{x}} \ln p_t(\mathbf{x})$ *on the tangent space of the data manifold* $T_{\pi(\mathbf{x})}\mathcal{M}$ *to the projection on the normal space* $\mathcal{N}_{\pi(\mathbf{x})}\mathcal{M}$ *approaches zero as* t *approaches zero, i.e.*

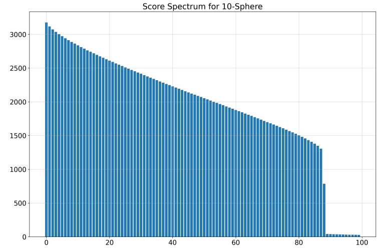
$$\frac{\|\mathrm{T}\nabla_{\boldsymbol{x}} \ln p_t(\mathbf{x})\|}{\|\mathrm{N}\nabla_{\boldsymbol{x}} \ln p_t(\mathbf{x})\|} \to 0, \ \textit{as } t \to 0.$$

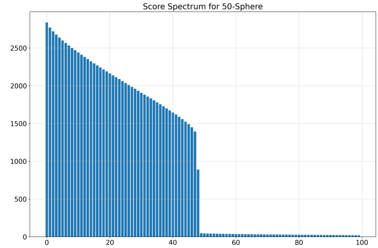
where N *and* T *are projection matrices on* $\mathcal{N}_{\pi(\mathbf{x})}\mathcal{M}$ *and* $T_{\pi(\mathbf{x})}\mathcal{M}$ *respectively. Therefore for sufficiently small* t *the score* $\nabla_{\mathbf{x}} \ln p_t(\mathbf{x})$ *is (effectively) contained in the normal space* $\mathcal{N}_{\pi(\mathbf{x})}\mathcal{M}$*.*

Experiments

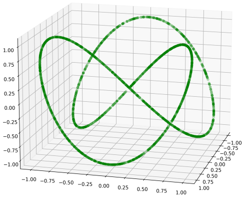
- k-spheres embedded in 100 dimensions with a random isometric embedding
- spaghetti line embedded in 100 dimensions
- synthetic image manifolds embedded in 1024 dimensions.
- MNIST

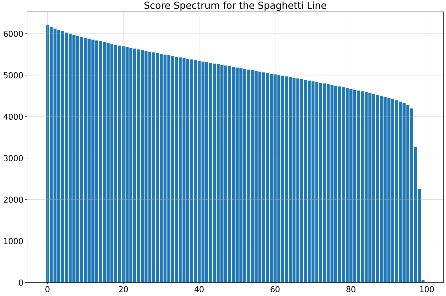
k-spheres in 100 dimensions



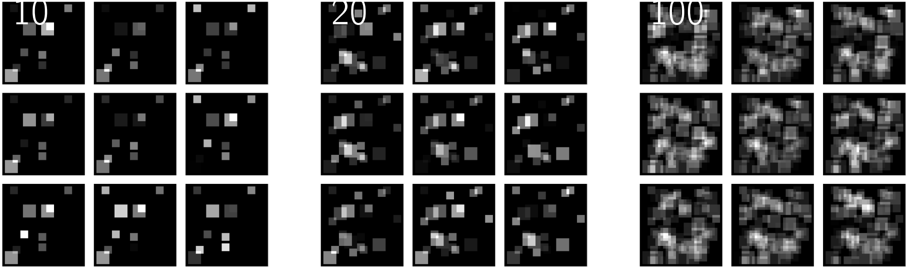


Spaghetti in 100 dimensions



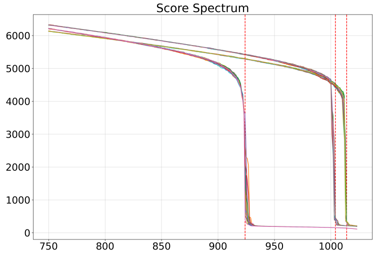


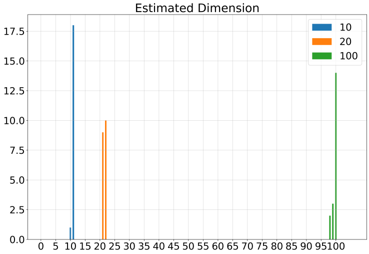
Square Manifold



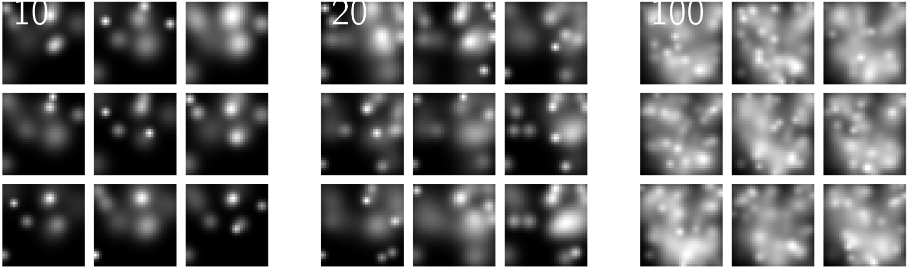
Samples from the Square Manifold for different dimensions.

Square Manifold



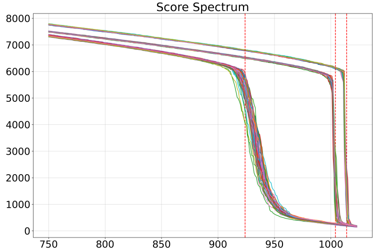


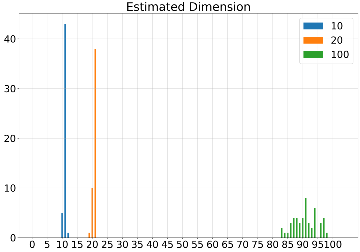
Gaussian Blobs Manifold

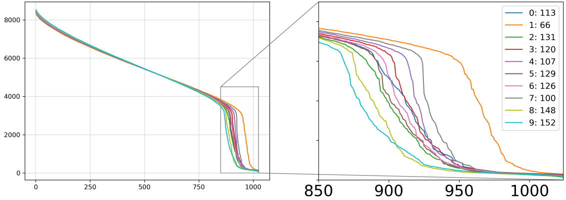


Samples from the Gaussian Blobs Manifold for different dimensions.

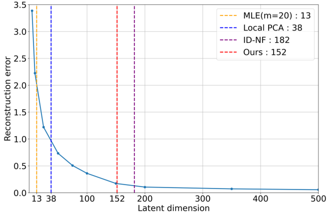
Gaussian Blobs Manifold







MNIST score spectra



Autoencoder validation

0	1	2	3	4	5	6	7	8	9
113	66	131	120	107	129	126	100	148	152

Table: Estimated dimension for each digit

Summary of experimental results

	Ground Truth	Ours	ID-NF	MLE (m=5)	Local PCA	PPCA
Euclidean Data Manifolds						
10-sphere	10	11	11	9.61	11	11
50-sphere	50	51	51	35.52	51	51
Spaghetti line	1	1	1	1.01	32	98
Image Manifolds						
Squares						
$k = 10$	10	11	9.7	8.48	10	10
$k = 20$	20	22	19.5	14.96	20	20
$k = 100$	100	100	94.2	37.69	78	99
Gaussian blobs						
$k = 10$	10	12	9.8	8.88	10	136
$k = 20$	20	21	17.8	16.34	20	264
$k = 100$	100	98	56.3	39.66	18	985
MNIST	N/A	152	182	14.12	38	706

Table: Comparison of dimensionality detection methods on various data manifolds.

Limitations

- *Approximation error*: Caused by imperfect score approximation $s_\theta(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \ln p_t(\mathbf{x})$.
- *Geometric error*: Arises when t isn't sufficiently small, leading to:
    - Increased tangential component of the score vector.
    - Differences in normal spaces across sampled points due to manifold curvature.

Conclusions

- Our estimator offers accurate ID estimates even for high dimensional manifolds, indicating superior statistical efficiency to statistical methods.
- This improvement is credited to the inductive biases of the unconstrained neural network (NN) estimating the score function, the critical quantity for ID estimation.
- Our theoretical results show that the diffusion model approximates the normal bundle of the manifold (more information than just the ID). We can potentially use a trained diffusion model to extract other important properties of the data manifold, e.g. curvature.