Score-based generative model learn manifold-like structures with constrained mixing

Li Kevin Wenliang
DeepMind
kevinliw@deepmind.com

Ben Moran
DeepMind
benmoran@deepmind.com

Abstract

How do score-based generative models (SBMs) learn the data distribution supported on a low-dimensional manifold? We investigate the score model of a trained SBM through its linear approximations and subspaces spanned by local feature vectors. During diffusion as the noise decreases, the local dimensionality increases and becomes more varied between different sample sequences. Importantly, we find that the learned vector field mixes samples by a non-conservative field within the manifold, although it denoises with normal projections as if there is an energy function in off-manifold directions. At each noise level, the subspace spanned by the local features overlap with an effective density function. These observations suggest that SBMs can flexibly mix samples with the learned score field while carefully maintaining a manifold-like structure of the data distribution.

1 Introduction

Score-based generative models (SBM) have achieved great success in learning natural data distributions [7, 8], which according to the manifold hypothesis are supported on low-dimensional manifolds. Unlike decoder-based methods such as VAEs and GANs, most SBMs do not explicitly construct distributions on low-dimensional manifolds with a compressed representation. Then, how do SBMs approximate distributions on manifold?

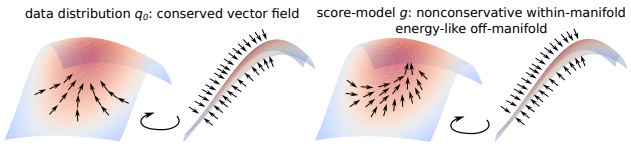

Figure 1: Schematic of our findings.

Given a dataset $\{x^{(i)}\}$ drawn i.i.d. from an unknown data density q_0 on $\mathbb{R}^d$, a score model $g_{\sigma_t}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ estimates the score function of the noisy data distribution $\nabla_x \log q_{\sigma_t}(x)$, where σ_t is the standard deviation of the additive Gaussian noise at time t, and $t \in [0, 1]$. Typically, as t decreases, σ_t decreases towards zero, and $q_1(x)$ is almost identical to a Gaussian. The estimated score function can then be used to define a stochastic or deterministic process to generate samples close to $q_0(x)$ by initialising from Gaussian samples. This procedure forms the basic methodology of SBM.

Fig. 1 summarises our main finding. The left two panels are two rotated views of the example data density q_0 (blue-orange) on a 2D manifold in a 3D ambient space. The density implies a conserved vector field. The right two panels show the vector field of a score model g which approximates the score of q_0 but is not guaranteed to be conservative. We find that, in this score model, the vector field is non-conservative only within the manifold; whereas the field in directions normal to the manifold remains close to the conservative score field of the noisy data distribution, constraining the samples to stay around the data manifold. Further, the local features of g span the same local subspace of an effective density function that is consistent with g in the sense we clarify soon.

NeurIPS 2022 Workshop on Score-Based Methods.

2 Local orthogonal features of approximate score functions

A local approximation of the score $g_\sigma(x) \approx g_\sigma(x_0) + \nabla_x g_\sigma(x)|_{x_0}(x - x_0)$ around x_0 involves the score Jacobian, so we use its singular value decomposition (SVD) to analyse g_σ locally:

$$\nabla_x g_\sigma(x) = \sum_{i=1}^{d} u_i(x) s_i(x) v_i^\intercal(x), \tag{1}$$

where the features $u_i(x), v_i(x)$ and $s_i(x) \geq 0$ are, respectively, the i'th left singular vector, right singular vector, and singular value of $\nabla_x g_\sigma(x)$, ordered so that $s_i < s_j$ for $i < j$. If g_σ is the score of a density function, then g_σ is conservative, $\nabla_x g_\sigma(x)$ is symmetric (or normal), and $u_i = \pm v_i$ for all i. To build more intuitions about score Jacobians, consider the multivariate Gaussian $\mathcal{N}(x; \mu, \Sigma)$ which has a constant score Jacobian $-\Sigma^{-1}$. Each pair of its singular vectors have *opposite signs*. In particular, the *singular value* of rank i is equal to the *inverse variance* along the i'th singular vector.

Now, suppose that the data distribution q_0 is a low-dimensional Gaussian supported on a subspace of $\mathbb{R}^d$. After adding a small additive Gaussian noise on $\mathbb{R}^d$, the SVD of the score Jacobian shows interpretable properties of the data distribution: large singular values or small variances appear along directions with abrupt changes in the score, reflecting steep curvatures along the off-manifold directions. Conversely, small singular values or large variances are associated with on-manifold directions along which the data density varies smoothly. This pattern generalises to curved manifolds as long as the noise is small compared to the local curvature. In practice, the Jacobian of a learned score estimator $\nabla_x g_\sigma(x)$ may not be symmetric as that of the Gaussian, but we can compare it to an effective conservative (energy-based) score field.

Proposition 1. *Given a model score $g_{\sigma_t}(x)$, there exists a conservative field $\tilde{g}_{\sigma_t}(x)$ such that for every path $x_{1:t}$ from g_{σ_t}, the likelihood of x_t computed under g_{σ_t} and $\tilde{g}_{\sigma_t}$ are identical.*

Proof. Suppressing the subscripts for brevity, we decompose ∇g as the sum of a symmetric component $\nabla \tilde{g} := 0.5(\nabla g + \nabla^\intercal g)$ and a skew-symmetric component $H := 0.5(\nabla g - \nabla^\intercal g)$. By Poincaré's Lemma, the symmetric $\nabla \tilde{g}$ implies the existence of a conservative field $\tilde{g}$. Given a sample path $x_{1:t}$, the likelihood of x_t depends on g through its divergence or the trace of the score Jacobian $\mathrm{Tr}(\nabla \tilde{g})$ [7, eqn. 39]. Then, we have $\mathrm{Tr}(\nabla g) = \mathrm{Tr}(\nabla \tilde{g} + H) = \mathrm{Tr}(\nabla \tilde{g}) + \mathrm{Tr}(H) = \mathrm{Tr}(\nabla \tilde{g}) + 0 = \mathrm{Tr}(\nabla \tilde{g})$. □

This means that, *given a sample path*, $\tilde{g}_\sigma$ is a valid score of an (unknown) density function equivalent to g_σ in terms of x_t's likelihood. Similar to but unlike the SVD of $\nabla_x g_\sigma(x)$, the eigendecomposition of the symmetric $\nabla_x \tilde{g}_\sigma(x)$ reveals the local features of the equivalent density: following the intuition of the Gaussian distributions, we see that the eigenvectors with negatively large eigenvalues correspond to off-manifold directions ($-\Sigma^{-1}$ is negative semi-definite); eigenvectors with small-in-magnitude eigenvalues indicate on-manifold directions. Positively large eigenvalues indicate positive curvature, and we find that they exhibit on-manifold features as shown in our experiments.

3 Experiments

We now empirically compare the SVD of $\nabla_x g_\sigma(x)$ and the eigendecomposition of $\nabla_x \tilde{g}_\sigma(x)$ to discover local geometries of a learned score model. For ease of visualisation, we used an SBM trained on MNIST digits [4], so $d = 28 \times 28 = 784$. The model is based on a U-net, adapted from the model of Song et al. [8]. More details are in Appendix A. For a given generated x at noise level σ, we denote the (minor) left and right singular vectors corresponding to the smallest singular value of $\nabla_x g_\sigma(x)$ as u_- and v_-, respectively, and the (major) vectors to the largest singular value as u_+ and v_+. For the "symmetrised" score $\nabla_x \tilde{g}_\sigma(x)$, we denote the minor and major eigenvectors as w_- and w_+, respectively, and the eigenvector with smallest-in-magnitude eigenvalue as w_0.

3.1 Local principal directions

Fig. 2 shows the generated samples at five time steps in the first column and their feature vectors in the other columns. Vectors v_-, u_- and w_0 reflect more general distortions of the strokes at higher noise levels, and more localised changes at lower noise levels. However, u_- and v_- are not well aligned, reflecting non-conservation of the score model, or a non-normal score Jacobian. The eigenvectors w_0 and w_+ of the symmetrised Jacobian show similar on-manifold semantics, but they are also not aligned with the singular vectors. Major vector u_+ and v_+ show high-frequency contents in the

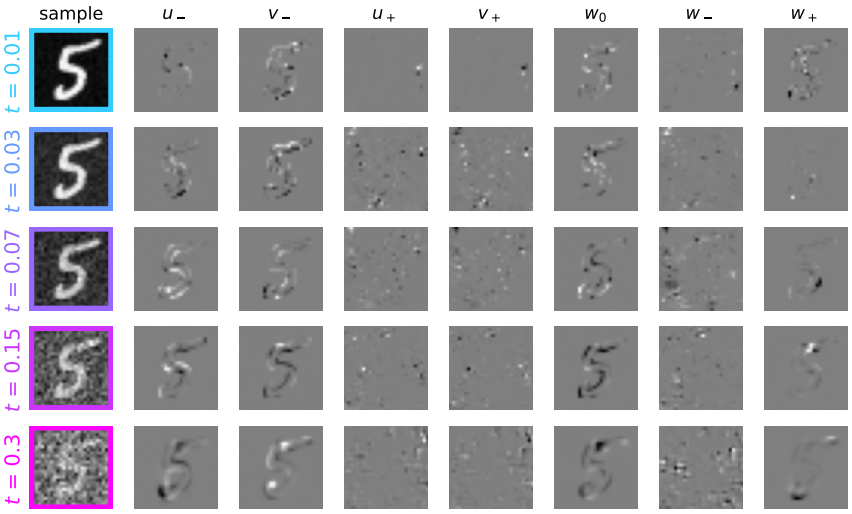

Figure 2: An example image sequence generated from the SBM and the singular and eigen-vectors.

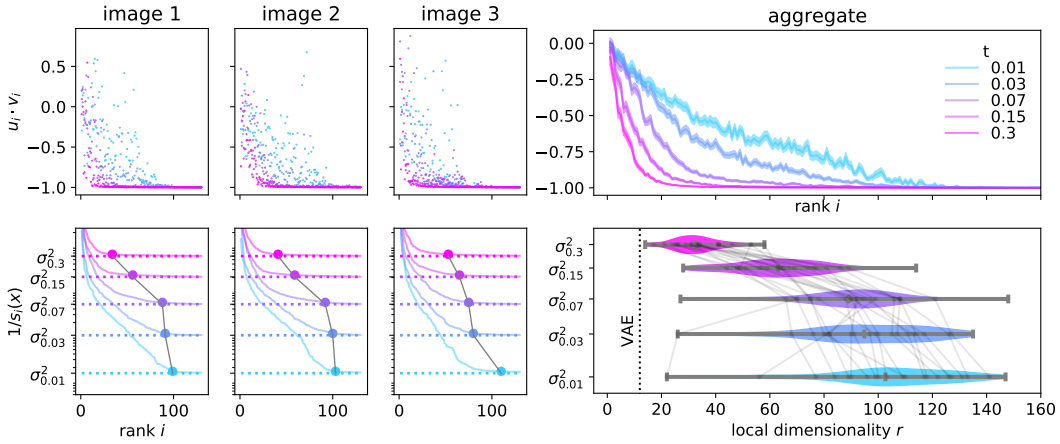

Figure 3: Top, cosine angle between the singular vectors. Bottom, singular values spectrum and effective local dimensionality corresponding to singular values above the noise floor.

background, which are likely off-manifold directions, as following these directions makes a sample less realistic. Unlike the minor singular vectors, these major vectors differ by a sign and points towards opposite directions. The eigenvector w_- of the effective energy-based field also shows high-frequency, off-manifold patterns in the background, similar to the major singular vectors.

In Fig. 3 (top), we show the inner products $u_i(x) \cdot v_i(x)$ (cosine angle) for three other example image sequences and the summary statistics of 128 images. They confirm that the singular vectors are not aligned for lower ranks and they differ by a sign for higher ranks. The opposite signs between u_i's and v_i's in higher ranks are consistent with the Gaussian score along off-manifold directions. But how to interpret the misalignment at lower ranks? The example image at $t = 0.3$ in Fig. 2 gives an interesting clue: its v_- focuses on the outline of the digit 5 but then maps to a u_- that resembles the outline of 6. This misalignment of singular vectors mixes data across different components, and we speculate that it could facilitate mixing between high-level image classes at high noise levels.

3.2 Singular values reveal a progressive and diverse manifold expansion

Following from our Gaussian intuition, the inverse singular values of a score Jacobian are related to the local length scales along principal directions of the distribution. We should then expect to see inverse singular values equal to the corresponding noise variances σ_t^2 along many potentially off-manifold directions. The examples in Fig. 3 (bottom left) support this prediction. In addition,

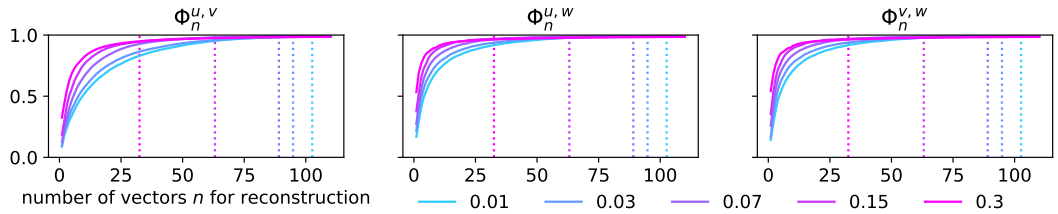

Figure 4: Solid lines: the amount of overlap between local subspaces. Dotted lines: mean number of nontrivial ranks for each temperature from Fig. 3 (bottom right).

the noise variance σ_t^2 lower bounds the inverse singular values, because the smoothing effect of the Gaussian noise determines the smallest length scale.

All nontrivial inverse singular vectors above the noise levels contribute to local data dimensionality, so we estimated the local data dimensionality of x_t by the smallest r_t such that $s_i^{-1}(x_t) < (1+\kappa)\sigma_t^2$ for all $i > r_t$, where we set $\kappa = 0.1$ as a slack constant. The coloured dots in Fig. 3 (bottom left) show the estimated local dimensionality for three images at each t, and Fig. 3 (bottom right) shows the distribution of thus estimated local dimensionalities for 128 images: they increases and spreads out more with decreasing noise levels. We compared this with the local dimensionality estimated by a VAE model [3, 2], described in Appendix B. The VAE-estimated local dimensionality of clean images turned out to be within 13 ± 3, substantially lower and less varied than the estimates from the score model in Fig. 3. Thus, these patterns of singular values suggest that the SBM has a much more expressive and more flexible representation power than the VAE.

3.3 Misaligned singular vectors span the same subspace and provides non-normal mixing.

From Figs. 2 and 3, we see that the singular vectors and eigenvectors are misaligned in the subspace given by nontrivial singular values. In principle, the right singular vectors define an input space of local changes, and the left singular vectors define an output basis to apply the denoising vector. To at least preserve mixing during diffusion, the left and right subspaces spanned by nontrivial u_i and v_i must substantially overlap, since otherwise the score field can project data into off-manifold directions. Further, they should also overlap with the eigen-subspace of the effective energy-based density spanned by $\{w_i\}$ with small-in-magnitude eigenvalues. We quantify the amount of overlap between these subspaces by how well one reconstructs the other for the first n vectors:

$$\Phi_n^{a,b} := 1 - \frac{1}{n}\sum_{j=1}^{n} \|\hat{a}_j^{b,n} - a_j\|_2^2 / \|a_j\|_2^2, \text{ where } \hat{a}_j^{b,n} = \sum_{k=1}^{n} (b_k \cdot a_j) b_k, \quad a, b \in \{u, v, w\}. \tag{2}$$

This is a criterion based on explained variance—a larger value indicates a better overlap. The results in Fig. 4 confirm a considerable overlap between the three subspaces within the nontrivial ranks at each t (Fig. 3). Therefore, although $g_\sigma(x)$ mixes image features onto different components in a non-normal fashion, it still updates the image within highly overlapped subspaces. The singular subspaces of $\nabla_x g_\sigma(x)$ also overlap well with the eigen-subspace of the symmetrised $\nabla_x \hat{g}_\sigma(x)$, suggesting that the learned singular subspaces conform to the data subspace of the effective density.

4 Conclusion

We have shown that the score field of a trained score-based model shows low local dimensionality consistent with the manifold hypothesis. As the noise scale decreases, the modelled manifold dimensionality rises and becomes more varied, reflecting increasingly precise description of samples. In addition to projecting noisy, off-manifold components back onto the manifold by normal projections, the score function mixes between feature directions within the data manifold using a non-conservative field, giving misaligned singular vectors. Nonetheless, these vectors span overlapping input and output subspaces, both of which also overlap with the local eigen-subspace of the effective energy-based field. Therefore, the score-based generative model learns a vector field that carefully constrains the samples to be within the manifold while mixing samples within the manifold. This explains the well-known harmless effect of non-conservation [6] and the little improvement obtained by encouraging conservation [1]. The discussion of a related work by Mohan et al. [5] is in Appendix C.

Acknowledgement We thank Conor Durkan, Mikołaj Bińkowski, Wendy Shang and Sander Dieleman for sharing their implementation of the diffusion model.

References

[1] Chen-Hao Chao et al. "Quasi-Conservative Score-based Generative Models". In: *arXiv preprint* (2022).

[2] Bin Dai, Li Wenliang, and David Wipf. "On the Value of Infinite Gradients in Variational Autoencoder Models". In: *Advances in Neural Information Processing Systems* (2021).

[3] Bin Dai and David Wipf. "Diagnosing and Enhancing VAE Models". In: *International Conference on Learning Representations*. 2018.

[4] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* (1998).

[5] Sreyas Mohan et al. "Robust And Interpretable Blind Image Denoising Via Bias-Free Convolutional Neural Networks". In: *International Conference on Learning Representations*. 2019.

[6] Tim Salimans and Jonathan Ho. "Should EBMs model the energy or the score?" In: *Energy Based Models Workshop-ICLR 2021*. 2021.

[7] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in Neural Information Processing Systems* 32 (2019).

[8] Yang Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*. 2020.

Score-based generative model learn manifold-like structures with constrained mixing: appendix

A The score-based model

The model, training objectives and procedures are essentially the same as the one proposed by [8].

A.1 Denoising networks

We took a standard implementation of the Unet and reduces the input size to $28 \times 28 \times 1$. The batch size is 2 048 and the learning rate is 0.0003. We trained the model for 540 000 steps.

A.2 Diffusion process

After obtaining the score estimates, we generated the images by discretising the SDE by the Euler-Maruyama method. The noise schedule is a monotically decreasing function of t, and we apply appropriate scaling on the drift term of the SDE (variance-preserving). We took samples at five noise levels at $t \in \{0.01, 0.03, 0.05, 0.15, 0.3\}$, approximately corresponding to noises with standard deviations $\sigma_t \in \{0.039, 0.10, 0.22, 0.46, 0.77\}$. For all summary statistics we used 128 randomly sampled images from the model so that they stay in the learned but approximate score field. Using more images does not change the conclusions.

B VAE model

Following [3, 2], we estimated the local dimensionality of a data point by the number of latent dimensions in which the posterior variances were close to zero. In practice, we regarded any posterior variance less than 0.5 as being close to zero. The VAE model we used to obtain a baseline data dimensionality has a standard Gaussian prior on $\mathbb{R}^{100}$. The likelihood is a Gaussian with its mean given by a neural network function of the latent variable and with variance as a free positive parameter. The variational posterior is a factorised Gaussian with means and variances as output of a neural network function. All neural networks have convolutional, ReLU nonlinearity and batch normalisation layers similar to the DCGAN.

We trained the network for 1 000 epochs on 60 000 MNIST digits. The estimated local dimensionality varies between digits: the digit "1" in general has lower dimensionality (around 11) than the other digit (around 18). This number decreases when we trained on noisy images, consistent with the effect of noise smoothing.

C Related work

Mohan et al. [5] used SVD to analyse locally linear features of the trained denoising model which was not used for diffusion-based generation. There, the trained model $f(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ outputs the denoised image under the expected l-2 loss directly rather than the score function. Also, the networks are constrained to be bias-free and have ReLU nonlinearities, so that the network is locally affine: $f(x) = \nabla f(x)x$.

We kept our score network to be as general as possible so did not impose these constraints. If we adopt these constraints, we will have $g(x) = \nabla g(x)x$. Under Gaussian noise with variance σ^2, when both f and g are perfectly trained, the optimally (in terms of expected l-2 loss) denoised estimate is

$$\hat{x} = x + \sigma^2 g(x) = x + \sigma^2 \nabla g(x)x = (\mathbf{I} + \sigma^2 \nabla g(x))x = \nabla f(x)x, \tag{3}$$

a form of the Stein's unbiased risk estimate. Clearly, the f used by Mohan et al. [5] and the g used by us are equivalent parametrisations of each other.

One should then expect to find the same conclusions between this paper and theirs. A seemingly contradictory finding is that, in Mohan et al. [5], the left and right singular vector of $\nabla f(x)$ at a given

rank were mostly aligned , whereas those of $\nabla g(x)$ here are mostly opposite to each other. This is, however, expected. In the idealised case where the score function is perfectly estimated, ∇g is symmetric, and so is ∇f. For small noise, the score Jacobian ∇g is symmetric with eigenvalues lower-bounded by $-1/\sigma^2$, since this is the lowest noise present in the data. Applying the tranformation between f and g in (3) shows that ∇f is in fact positive semidefinite, which has identical left and right singular vectors. However, Mohan et al. [5] did not interpret the misaligned singular vectors, which we explain as arising from the non-conservation of the approximate score model.