Heavy-tailed denoising score matching

Jacob Deasy [1] **Nikola Simidjievski** [1] **Pietro Liò** [1]

Abstract

Score-based model research in the last few years has produced state of the art generative models by employing Gaussian denoising score-matching (DSM). However, the Gaussian noise assumption has several high-dimensional limitations, motivating a more concrete route toward even higher dimension PDF estimation in future. We outline this limitation, before extending the theory to a broader family of noising distributions—namely, the generalised normal distribution. To theoretically ground this, we relax a key assumption in (denoising) score matching theory, demonstrating that distributions which are differentiable *almost everywhere* permit the same objective simplification as Gaussians. For noise vector length distributions, we demonstrate favourable concentration of measure in the high-dimensional spaces prevalent in deep learning. In the process, we uncover a skewed noise vector length distribution and develop an iterative noise scaling algorithm to consistently initialise the multiple levels of noise in annealed Langevin dynamics. On the practical side, our use of heavy-tailed DSM leads to improved score estimation, controllable sampling convergence, and more balanced unconditional generative performance for imbalanced datasets.

1. Introduction

Given a probability distribution $p(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, the *score function* is defined as

$$s(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}), \tag{1}$$

the gradient of the log-density with respect to the input $\mathbf{x}$. The score is a vector field of the gradient at $\mathbf{x}$, and gives the direction of the maximum increase in log-density.

Score based models (SBMs) are parameterised and trained to estimate $\nabla_{\mathbf{x}} \log p(\mathbf{x})$. Unlike likelihood-based models, such as normalising flows (Rezende & Mohamed, 2015; Kobyzev et al., 2020) or autoregressive models (Papamakarios et al., 2017), this approach has the advantage of modelling an unconstrained function that does not need to be normalised.

By starting with the energy based model formulation

$$p_\theta(\mathbf{x}) = e^{-f_\theta(\mathbf{x})}/Z_\theta, \tag{2}$$

for parameters $\theta \in \mathbb{R}^m$, with $m \gg 1$ for deep learning models, it is clear that

$$s_\theta(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}) = -\nabla_{\mathbf{x}} f_\theta(\mathbf{x}) - \underbrace{\nabla_{\mathbf{x}} \log Z_\theta}_{=0} = -\nabla_{\mathbf{x}} f_\theta(\mathbf{x}), \tag{3}$$

naturally removes the oft intractable *partition function* Z_θ.

1.1. Score matching

The goal of SBMs is to fit $s_\theta(\mathbf{x}) := \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$ to $\nabla_{\mathbf{x}} \log p_{\mathbf{x}}(x)$, but of course $\nabla_{\mathbf{x}} \log p_{\mathbf{x}}(\mathbf{x})$ is not available in the first place. As such, it is necessary to assess whether any given minimisation can avoid the tautologous use of $p(\mathbf{x})$.

A simple first attempt to minimise the Euclidean distance, known as the Fisher divergence, across the space gives the explicit score matching (ESM) objective

$$\mathcal{J}_{ESM_p}(\theta) = \frac{1}{2}\mathbb{E}_{p(\mathbf{x})}\left[||\nabla_{\mathbf{x}} \log p(\mathbf{x}) - s_\theta(\mathbf{x})||_2^2\right]. \tag{4}$$

Despite the continuing presence of $p(\mathbf{x})$, a useful result is that, following an integration by parts, $\mathcal{J}_{ESM_p}$ (ignoring a constant shift) simplifies to implicit score matching (ISM)

$$\mathcal{J}_{ISM_p}(\theta) = \mathbb{E}_{p(\mathbf{x})}\left[\frac{1}{2}\|s_\theta(\mathbf{x})\|_2^2 + \text{tr}\left(\nabla_{\mathbf{x}} s_\theta(\mathbf{x})\right)\right], \tag{5}$$

where *the density function of the observed data does not appear* (Hyvärinen, 2005). This integration is subject to a few weak constraints which are detailed here as they motivate a theorem in Section 2.

#1 The PDF $p(\mathbf{x})$ is differentiable.

#2 $\mathbb{E}_{p(x)}\left[\left\|\frac{\partial \log p(\mathbf{x})}{\partial \mathbf{x}}\right\|^2\right]$ is finite.

[1]Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom. Correspondence to: Jacob Deasy <jd645@cam.ac.uk>.

#3 For any θ:

A $\mathbb{E}_{p(\mathbf{x})}\left[\|s_\theta(\mathbf{x})\|^2\right]$ is finite.
B $\lim_{\|\mathbf{x}\|\to\infty}[p(\mathbf{x})s_\theta(\mathbf{x})] = 0$.

In practice, discretising the expectation, $\mathcal{J}_{ISMp}$ is then approximated by

$$\mathcal{J}_{ISMp_0}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}\left\|s_\theta\left(\mathbf{x}^{(i)}\right)\right\|_2^2 + \text{tr}\left(\nabla_{\mathbf{x}} s_\theta\left(\mathbf{x}^{(i)}\right)\right)\right], \tag{6}$$

for N data samples, an intuitive objective where:

- Term one minimises the scale of the score to zero, inducing the presence of a local minimum or maximum.
- Term two, the trace of the Jacobian of the score, being minimised then clearly indicates an objective forcing local maxima at each data point.

1.2. Denoising score matching

Problematically, the trace of the Jacobian in (5) and (6) requires $\mathcal{O}(n)$ backpropagations to calculate and is therefore computationally expensive enough to render this objective impractical. As an example of suggested optimisations, Song et al. (2020a) proposed *sliced score matching* (SSM) which projects the vectors onto random directions (far fewer than n times) and takes the expectation of the objective over these directions. The sliced Fisher divergence is then approximated by

$$\frac{1}{2}\mathbb{E}_{p_\nu}\mathbb{E}_{p_{\text{data}}}\left[\left(\nu^T\nabla_{\mathbf{x}}\log p_\theta(\mathbf{x}) + \nu^T\nabla_{\mathbf{x}}\log p_{\text{data}}(\mathbf{x})\right)^2\right], \tag{7}$$

where ν is a random direction based on a given distribution p_ν. However, SSM has recently been superseded by a new form of *denoising score matching* (DSM), originally from (Vincent, 2011), which avoids the Jacobian altogether.

The first step of DSM is to perturb the data $\mathbf{x}$ with a known noise distribution $q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})$ (normally convolution with a diagonal multivariate Gaussian kernel)

$$q_\sigma(\mathbf{x}) = \int_{\mathbf{X}} q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})p_{\text{data}}(\mathbf{x})\text{d}\mathbf{x}. \tag{8}$$

The key step in (Vincent, 2011), relying on the same assumptions in a similar integration by parts to that of (Hyvärinen, 2005), was to prove that (5) is equivalent (as an objective, i.e. up to fixed constants) to DSM

$$\mathcal{J}_{DSMq_\sigma}(\theta) = \frac{1}{2}\mathbb{E}_{q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})p_{\text{data}}(\mathbf{x})}\left[\|s_\theta(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}}\log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})\|_2^2\right], \tag{9}$$

with $s_{\theta^*}(\mathbf{x}) = \nabla_{\mathbf{x}}\log q_\sigma(\mathbf{x})$ almost surely, and $\nabla_{\mathbf{x}}\log q_\sigma(\mathbf{x}) \approx \nabla_{\mathbf{x}}\log p_{\text{data}}(\mathbf{x})$ when the noise is low enough for $q_\sigma(\mathbf{x}) \approx p_{\text{data}}(\mathbf{x})$. Crucially, perturbation of the distribution in (9) is computationally trivial and only a single backpropagation is required.

Taking stock, a lot has been achieved toward making likelihood estimation more tractable. For the entirely generalisable formulation of the likelihood as an EBM: the partition function has been discarded, the model likelihood has been integrated out, and a tractable equivalent of an intuitive objective was obtained. Nevertheless, important questions remain about how a well-approximated score should be best-used and how samples can be generated.

1.3. Langevin dynamics

The answer to the generation problem is to sample by using score estimates to ascend the gradient for a given input. Due to the monotonic nature of the logarithm function, iteratively following the direction of the largest score estimate is equivalent to performing gradient ascent on the data distribution. As such, multiple iterative optimisation algorithms are available for use at this stage.

The procedure of choice, Langevin Monte Carlo (LMC, (Besag, 1994)), is a Markov Chain Monte Carlo (MCMC) method for obtaining random samples from probability distributions for which direct sampling is difficult. The goal is to follow the gradient but add a bit of noise so as to not get stuck at local optima, explore the entire distribution, and sample from it.

The LMC sampling procedure for $p(\mathbf{x})$, using only $\nabla_{\mathbf{x}}\log p(\mathbf{x})$ and remaining uncorrected as Langevin Dynamics (LD), is summarised in Algorithm 1. The major assumptions of note here are the standard choice of Gaussian noise in LD, which can be replaced with heavier tailed noise sources (Şimşekli, 2017), as well as the first-order nature of the iteration. In the last few years, the convergence rates of higher-order schemes of LD have been formalised (Cheng et al., 2018; Mou et al., 2019) and their integration with score matching looks to be an interesting avenue of research beyond the scope of this document.

Algorithm 1 Langevin dynamics.

Input Hyperparameters of prior $\pi(\mathbf{x})$, step size $\varepsilon \ll 1$, step limit T, and initialised $\tilde{\mathbf{x}}_0 \sim \pi(\mathbf{x})$.
for $t = 1 \dots T$ **do**
  Sample:
  $\mathbf{z}_t \sim \mathcal{N}(0, I)$ $\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1} + \varepsilon\nabla_{\mathbf{x}}\log p\left(\tilde{\mathbf{x}}_{t-1}\right) + \sqrt{2\varepsilon}\mathbf{z}_t$
end for

1.4. Limitations and recent improvements

Now that an end-to-end solution to sampling—score estimation through DSM followed by LMC—has been established, it is apt to consider this process' inherent limitations:

1. *The manifold hypothesis*. High dimensional data tend to lie in low dimensional manifolds, so $\nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})$ is undefined in some regions. This hypothesis is central to the majority of NN theory and DL architectures. As such, it is potentially both computationally wasteful and mathematically fruitless to estimate a score along certain of the input space's dimensions.

2. *Inaccurate score estimation in low data-density regions*. The reverse statement of the manifold hypothesis states that data will not lie in the majority of the input space. Therefore, vector field estimation can be misled around low density regions where there are too few samples to guide the correct direction of the score vector, even if the dimension is, overall, relevant.

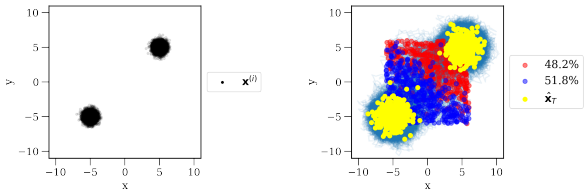
3. *Slow mixing of Langevin Dynamics between data modes*. With disconnected distribution support, or a mixture of two disjoint components with a weighting coefficient π, scores cannot recover this coefficient and are invariant towards mode weights (Song & Ermon, 2019). For instance, a mixture of Gaussians with very different means and low variances will be difficult to fully model due to the light tails of the Gaussian diffusion in LMC—it is very unlikely that a transition between modes will take place.

To address these issues, Song & Ermon (2019) suggested *annealed Langevin dynamics* (ALD, see Algorithm 2). Follow-up work then made five training technique suggestions which allow improved scale and generation quality (Song & Ermon, 2020). The success of ALD has been such that recent SBMs are now on par, if not better (with heavy compute), than best-in-class GANs and autoregressive models (Song et al., 2020b; Vahdat et al., 2021).

Algorithm 2 Annealed Langevin dynamics.

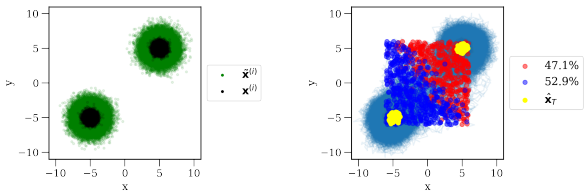
Input Gaussian noise scaling factors $\{\sigma_1, \ldots, \sigma_k \in \mathbb{R}_+ \; s.t. \; \sigma_1 > \ldots > \sigma_k\}$, and parameters to run LD.
for $i = 2 \ldots k$ **do**
  Run LD_i with noise level σ_i starting from the result of LD_{i-1}
end for

Due to the success of these improvements, research in this area has proliferated over the past two years. As a full review is beyond the scope of this work and yet to appear in the literature, a brief summary is included here. Critiques and expansions of both discrete and continuous (see Section A.3) DSM have been presented in (Huang et al., 2021; Kim et al., 2021; Song et al., 2021). DSM for discrete data was formally defined in (Hoogeboom et al., 2021), and techniques for sampling with score (and higher order (Meng et al., 2021)) estimates have experienced a renaissance (Jolicoeur-Martineau et al., 2021). Additionally, the connection between SBMs and denoising diffusion probabilistic models (DDPMs) has been clarified (Ho et al., 2020; Song et al., 2020b). Finally, closely related to the next section, is the first use of non-Gaussian noise in DDPMs in Nachmani et al. (2021), to assess the effects of noise with more degrees of freedom.



(a) $p(\mathbf{x})$ (b) Sample paths.

Figure 1. DSM training and LD sampling. In **a**, $p(\mathbf{x})$ is modelled as an additive mixture of ($k = 2$) bivariate Gaussians with 20,000 samples. An MLP is trained to estimate the score from samples noised by $q_\sigma(\tilde{\mathbf{x}}|\mathbf{x}) \sim \mathcal{N}(\mathbf{x}, \mathbf{I})$. 1,000 sampled paths are evolved in **b** to demonstrate the decision boundary, its asymmetry (relevant for class imbalance), and the upper bound on approximation accuracy due to the underlying unit noise. Full details in Figure 11.



(a) (b)

Figure 2. Multiple noise level DSM training and ALD sampling. The setup and figures are identical to Figure 1 except that two noise scales, $\sigma_1 = 1.0$ and $\sigma_2 = 0.25$, are used. Full details in Figure 12.

2. High dimensional noising

The previous section discussed the background of denoising score matching and finished with various strategies to scale this process to higher dimensions and better sample quality. This section will consider the weaknesses of those scaling strategies, with a particular focus on generalising high dimensional noise perturbations.

2.1. Beyond Gaussian noise

According to the derivation in Vincent (2011), the choice of Gaussian noise is simply for convenience and has the bonus of an intuitive score

$$\nabla_{\tilde{\mathbf{x}}} \log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x}) = \Sigma^{-1}(\mathbf{x} - \tilde{\mathbf{x}}), \tag{10}$$

which corresponds to moving from noisy $\tilde{\mathbf{x}}$ to clean $\mathbf{x}$.

To elucidate how Gaussian DSM works in practice, in Figure 1, a $2D$ example is provided, demonstrating the method converging and generating samples from a mixture of Gaussians. The example considers all steps of the procedure: noising, training, and sampling. Then, in Figure 2, the example is extended to the multiple noise levels in DSM with ALD. Although the improvement between the two is evident, this synthetic example will be used to highlight the weaknesses of *Gaussian* DSM with ALD in Subsection 3.1.

By considering the actual constraint on the noise distribution, that $\log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})$ is differentiable (Vincent, 2011), this subsection will explore the consequences of the pivotal Gaussian noise assumption in $\mathcal{J}_{DSMq_\sigma(\theta)}$. The differentiability condition encompasses a broad range of potential distributions and gives rise to the questions: *What form could and should q_σ take? How does the choice of q_σ influence model learning?*

2.1.1. Gaussian noise in high dimensions

Now that intuition about the role of the noising process in DSM has been established, it is necessary to consider the effects of noising in higher dimensions. Deep learning excels at minimising function approximation error in high dimensional space (Barron, 1994). However, intuition about high-dimensional statistics is infamously poor (Bishop et al., 1995) and, as they have evolved far beyond low-dimensional statistics, canonical 'small' DL datasets appear sparse when considered in their full domain. In the case of images, common generative baselines range from MNIST with samples in $\mathbb{R}^{784}$ (LeCun et al., 2010), to Flickr Faces HQ (FFHQ) with samples in $\mathbb{R}^{3,145,728}$ (Karras et al., 2019). For time series with a discrete parameterisation, samples inhabit $\mathbb{R}^{T\times d}$, where d can be particularly high in medical datasets. Moreover, all of these high dimensions precede any considerations about multimodality, where several high-dimensional input domains (or their compressed representations) can be fused using the flexibility of neural architectures.

For instance, consider how the *squared L^2 length distribution* of the isotropic Gaussian vector

$$Y = \|\mathbf{X}\|_2^2, \tag{11}$$

follows either of the chi-squared distributions

$$Y \sim \chi^2(n) = n\chi^2(1), \tag{12}$$

which approaches a Gaussian distribution centred at n in the limit $n \to \infty$. See Section A.1 for the derivation, Section 2.1.4 for a full description of the moments, and Figure 15 for a visualisation of the chi-squared distribution for increasing degrees of freedom.

2.1.2. The problem in high-dimensional SBMs.

As described in Section 1, the conditional DSM Gaussian noise distribution q_σ smooths around each data point $\mathbf{x}$. In the ideal scenario, the surrounding Gaussian n-spheres would overlap slightly, filling the high-dimensional convex hull defined by the dataset. As such, throughout the hull, the SBM would learn to faithfully interpolate the space, estimating gradients accurately so that they can be used in an iterative generation procedure. However, using standard results, it is now clear that in the high-dimensional setting of deep learning, these n-spheres in fact approach n-annuli—very thin shells.

This consideration immediately offers a new interpretation of why multiple levels of noise in ALD (see Section 1.4) were a major improvement over prior methods. Although the original motivation in (Song & Ermon, 2019) was to enable LD noise annealing, similar to annealed importance sampling (Neal, 2001), this step also stacked concentric noise annuli. Therefore, SBMs with ALD learn gradients that apply to a larger volume of the dataset interior. Moreover, for the original ALD paper, this perspective confounds the performance improvement due to annealing the Langevin dynamics with 'filling the convex hull'. Such an insight motivates decoupling of, and clarification around, the effect of both approaches.

Despite the follow-up improvements to ALD (Song & Ermon, 2020), recognising this concentration of noise and increasing the number of noise levels, the authors' motivation was to correctly balance coverage across regions of different weight. This interpretation can be taken further and permits several opportunities:

- Even with multiple levels of noise, how do these models fair when generating sparse distributions—what is the performance-sparsity trade-off? As highlighted in Figure 16, the Gaussian distribution has relatively light tails compared to several reasonably well-behaved distributions that have been studied in-depth. Intuitively, heavier tails should facilitate sampling further across sparse domains and aid score interpolation.
- At the time of writing, noise level selection for the best discrete ALD model is sampled linearly in log space between two hyperparameters for the minimum and maximum noise level. In the continuous case, recent models have tried to learn this distribution (Kingma et al., 2021), but the resulting approximation has not

been theoretically explained. Clearly, refinement of the discrete case, potentially leading to an explanation in the continuous case, is a motivating theoretical goal. Moreover, the relationships between data dimension, DSM noise distribution, and DSM noise length distribution have not been explored. In particular, Section A.2 addresses skewed length distributions and general noise with a quantile matching algorithm.

- Finally, the variance of the squared length distribution in (Song & Ermon, 2020) is incorrect. Unfortunately, this means that the subsequent derivations (Proposition 2 and 3, corresponding to technique 3 and 4, in (Song & Ermon, 2020)) are also invalid. The correct variance is detailed in (25) alongside the consequent differences for SBMs due to its new form.

As a result, it is tempting to turn toward common heavy tailed distributions, such as those in Figure 16. Unfortunately, as a first port of call, the Cauchy distribution is notoriously difficult to manipulate, evidenced by its undefined moments, and does not permit the same concentration analysis as in Section 2.1.1 (Eicker, 1985). The same issue, the *summation* of the squared RV rather than the squaring itself, arises for the Student-t distribution because the square of a t-distribution is an F-distribution (Fisher-Snedecor distribution (Box & Tiao, 2011)) which has an undefined MGF.

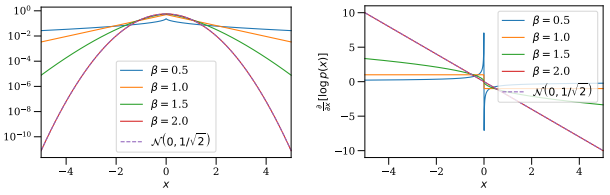
Finally, the Laplace distribution seems to be the only viable heavy-tailed distribution for a similar concentration of measure calculation to the Gaussian noise vector. The result is that the square of a Laplace RV follows a Weibull distribution and, as the Weibull distribution is linear in its first parameter, the squared length distribution is a Weibull distribution also. Notwithstanding the potential of this result, it is possible to go further by considering a much broader family of distributions that both subsumes the Gaussian and Laplace distributions and is similarly equipped with a tractable PDF.

2.1.3. Generalising to the generalised normal (exponential power) distribution

The generalised normal (GN) distribution (Nadarajah, 2005), $X_i \sim \mathcal{GN}(\mu, \alpha, \beta)$ with $\mu \in \mathbb{R}$ and $\alpha, \beta \in \mathbb{R}_+$, has PDF

$$f_X(x; \mu, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{-\left(\frac{|x-\mu|}{\alpha}\right)^\beta\right\}, \tag{13}$$

which recovers the standard Gaussian distribution for $(\alpha, \beta) = \left(\sqrt{2}, 2\right)$, the standard Laplace distribution for $(\alpha, \beta) = (1, 1)$, and the uniform density for $\beta = 0$. The GN is used when the concentration of values around the mean and the tail behaviour are of particular interest (Box & Tiao, 2011), apt for this case. The corresponding score of $1D$ GN noise is



(a) Log PDF. (b) Score function.

Figure 3. The generalised normal distribution for varied β.

$$\frac{d}{d\tilde{x}}\left[\log q_{\mathcal{GN}}(\tilde{x}|x)\right] = \frac{d}{d\tilde{x}}\left[-\left(\frac{|\tilde{x}-x|}{\alpha}\right)^\beta\right] \tag{14}$$

$$= -\frac{\beta}{\alpha^\beta}\text{sign}(\tilde{x}-x)|\tilde{x}-x|^{\beta-1} \tag{15}$$

(15) clearly indicates that the score of the generalised normal distribution is, however, continuous but not differentiable at zero. This contravenes the necessary assumption for DSM which was mentioned in Section 1.2 and (Vincent, 2011). Therefore, to use this more general family of generalised normal distribution noise, it is necessary to weaken the theoretical constraints to piecewise-differentiable distributions.

Theorem 2.1. *Assume that the estimated score function $s_\theta(\mathbf{x})$ obeys the assumptions outlined in Section 1.1, except that $s_\theta(\mathbf{x})$ is instead differentiable almost everywhere. Then, the objective function for $\mathcal{J}_{ESM_p}$ in (4) is still equivalent to $\mathcal{J}_{ISM_p}$ in (5). Proof in Appendix A.4.*

This theorem establishes that the generalised normal distribution, or similar distributions such as the Laplace distribution, can be used for noising in DSM. Therefore, to motivate its usage, further theoretical results are now derived surrounding GN concentration of measure.

2.1.4. Comparison of concentration moments

Similar to the Gaussian case, considering the squared L^2 length distribution Y, but deriving the distribution of $Z_i = X_i^2$ first for simplicity

$$F_{Z_i}(z) = P(Z_i \leq z) = P\left(X_i^2 \leq z\right) = P\left(|X_i| \leq \sqrt{z}\right), \tag{16}$$

and repeating the process in Section 2.1.1 gives

$$f_{Z_i}(z) = F'_{Z_i}(z) = \frac{1}{\sqrt{z}}\phi_{\mathcal{GN}}\left(\sqrt{z}\right) \tag{17}$$

$$= \frac{1}{\sqrt{z}}\frac{\beta}{2\Gamma(1/\beta)}\exp\left\{-|\sqrt{z}|^\beta\right\}, \tag{18}$$

where $\phi_{\mathcal{GN}}$ is the corresponding unit distribution $(\alpha, \mu) = (1, 0)$ and, as the square of the domain of X_i is $\mathbb{R}_+$, the modulus can be ignored.

Recalling the *generalised Gamma distribution* (Stacy, 1962), $\mathcal{GG}(a, d, p)$ with PDF

$$f(x; a, d, p) = \frac{p/a^d}{\Gamma(d/p)} x^{d-1} \exp\left\{-\left(\frac{x}{a}\right)^p\right\}, \tag{19}$$

with $d \in \mathbb{R}$ and $a, p \in \mathbb{R}_+$, it becomes clear that $Z_i \sim \mathcal{GG}(a = 1, d = 1/2, p = \beta/2)$ and the sum remains. Therefore

$$\begin{aligned} f_Y(y) &= \frac{1}{n}\phi_{\mathcal{GG}}\left(\frac{y}{n}\right) && (20)\\ &= \frac{1}{n}\frac{p}{a^d\Gamma(d/p)}\left(\frac{y}{n}\right)^{d-1} \exp\left\{-\left(\frac{y}{an}\right)^p\right\} && (21)\\ &= \frac{p}{(an)^d\Gamma(d/p)} y^{d-1} \exp\left\{-\left(\frac{y}{an}\right)^p\right\}, && (22)\end{aligned}$$

shows $n\mathcal{GG}(a, d, p) = \mathcal{GG}(na, d, p)$, which means

$$Y \sim \mathcal{GG}(a = n, d = 1/2, p = \beta/2). \tag{23}$$

This derivation establishes the length distribution appropriate for the generalised normal noise vector. To determine whether such a generalisation is useful, it is relevant to analyse how the moments of this distribution evolve with n.

For a Gaussian noise vector, the moments of $\|\mathbf{X}\|_2^2 = Y \sim \chi^2(n)$ are:

$$\begin{aligned} \mathbb{E}[Y] &= n && (24)\\ \text{Var}(Y) &= 2n && (25)\\ \text{Skew}(Y) &= \sqrt{\frac{8}{n}} \xrightarrow{n\to\infty} 0 && (26)\\ \text{Kurtosis}(Y) &= \frac{12}{n} \xrightarrow{n\to\infty} 0. && (27)\end{aligned}$$

The main properties of interest here are:

- Variance scales linearly with dimension. This moment dictates how thick the concentric annuli are for the sequence of noise levels in DSM with ALD.
- Skew and kurtosis tend to zero as dimensionality increases, leading to a near-Gaussian distribution in the limit. Also, when the skew value is non-zero, any scheme to overlap concentric annuli will be consistently biased toward one side of each annulus.

On the other hand, for the generalised Gaussian noise vector, the first two moments of $\|\mathbf{X}\|_2^2 = Y \sim \mathcal{GG}(a = n, d = 1/2, p = \beta/2)$ are:

$$\mathbb{E}[Y] = a\frac{\Gamma((d+1)/p)}{\Gamma(d/p)} = nC_1 \tag{28}$$

$$\text{Var}(Y) = a^2\left(\frac{\Gamma((d+2)/p)}{\Gamma(d/p)} - \left(\frac{\Gamma((d+1)/p)}{\Gamma(d/p)}\right)^2\right) = n^2C_2, \tag{29}$$

where

$$C_1 := \frac{\Gamma(3/\beta)}{\Gamma(1/\beta)} \tag{30}$$

$$C_2 := \frac{\Gamma(5/\beta)}{\Gamma(1/\beta)} - \left(\frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}\right)^2. \tag{31}$$

Relative to the previous moments, it should firstly be noted that (30) scales the mean in a nonlinear fashion. As depicted in Figure 18(a), (30) undergoes a super-exponential decay with respect to β. Therefore, use of low β noising strategies will push annulus samples relatively far from the base data point compared to Gaussian noise. It is apparent that a trade-off has emerged, between the desire for heavy-tails in DSM to fill high-dimensional space, and the unwieldy resulting length distributions. This is perhaps, unsurprising, given the renowned difficulties when working with Lévy-like distributions (Mandelbrot & Mandelbrot, 1982).

Secondly, the property that variance scales quadratically with dimension has surfaced. Therefore, as long as (31) is greater than $1/n$—almost a guarantee given that $n \gg 1$ and the even more aggressive exponential for low β in Figure 18(b)—substantially thicker shells will be present. Of course, this gain comes with the caveat that low β noise is likely to be problematic, corresponding to score functions with a singularity at zero (evidenced in Figure 3).

Despite the unintuitive form of the gamma functions comprising (30) and (31), both terms simplify for $\beta \in \{2, 1, \ldots, 1/k\}$, $k \in \mathbb{N}$, to Gaussian, Laplace, and closed-form moments respectively.

3. Results

After the theoretical groundwork of the previous section, this section designs empirical experiments to explore and confirm the utility of heavy-tailed denoising score matching (HTDSM). A qualitative and quantitative assessment of the novel insights of Section 2 is provided at multiple scales, each lending support to the use of HTDSM in practice.

3.1. Low dimensional space

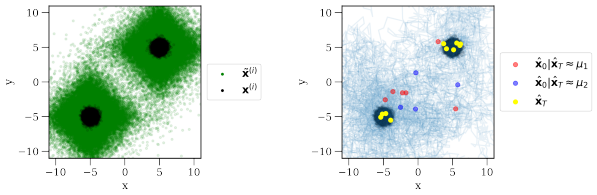
Before progressing to high-dimension DL image datasets, it is apt to begin with an easily controlled and visualised continuation of the $2D$ example given in Figures 1 and 2.[1]

[1]Code is available at github.com/jacobdeasy/heavy-tail-dsm.

As a first implementation of the HTDSM scheme described in Section 2, Figure 4 (expanded in Figure 13) combats the density approximation task of Figure 2 using Laplace ($\beta = 1$) noise. Figure 13(a) illustrates the diamond, rather than circular, noise structure of a diagonal bivariate Laplace distribution. Figure 13(b) and 13(c) respectively demonstrate that ALD training *and sampling* converge with Laplace (sub-Gaussian, piece-wise differentiable) noise, confirming Theorem 2.1. The effect of the heavier-tailed noise when sampling is evidently present for the first half of the noise levels in Figure 13(d), but this effect is outweighed by the down-scaling of ALD in the second half of sampling. Paths in Figure 13(e) begin from any point in the initialisation, extend across a far broader space, and all converge. The use of sub-Gaussian sampling diffusion is a novel step beyond standard ALD using SBMs and is closely aligned with fractional Langevin Monte Carlo methods (Şimşekli, 2017). A positive is the removal of any kind of decision boundary, but a negative is the slightly inaccurate final solution. One way to solve this inaccuracy would be to simply add another, lower noise, level of ALD. Figure 4 also clarifies that the higher variance in shell radii, derived in Section 2.1.4, is in fact the variance arising due to the non-spherical nature of the high-dimensional generalised normal distribution. For instance, in the case $\beta = 1$, Laplace noise provides samples in an approximate hypercube around its centre. The corners of the hypercube extend further along the axes than the Gaussian case, sacrificing probability mass not aligned with the coordinate system. It is noteworthy that this hypercube is approximate and the infinite domain of the Laplace distribution is therefore still more useful than the fixed hypercube of the uniform distribution. Immediate extensions are available, such as using a radial basis for the noise distribution, similar to that used in Farquhar et al. (2020) for Bayesian neural network parameterisation. However, this direction is beyond the scope of this work and the Cartesian basis will continue to be used throughout.

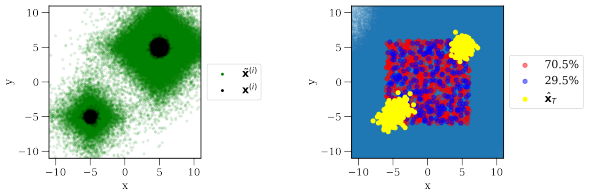
Figure 14 also depicts how standard DSM with ALD can suffer from mode collapse. The setup and subfigures are identical to Figure 2, except that $p(\mathbf{x})$ samples now have an imbalance of 10:1 between modes 1 (upper right) and 2 (lower left) respectively. Particle paths in Figure 14(b) and Figure 14(d) clearly show a preference for mode 1, even crossing mode 2 entirely. This is arguably not a problem, but 97.2% of particles approaching mode 1 in Figure 14(d) does not reflect the true imbalance ($> 99\%$ is also not an uncommon steady-state for this setup). The large scores associated with distant particle migration, across mode 2 to mode 1, can also be seen by the scale of the initial scores in Figure 14(c).

In addition, Figure 5 illustrates Laplace DSM with ALD for the class imbalance problem of Figure 14. 5(b) establishes that Laplace noise can compensate for class imbalance. In particular, 29.5% of particles finishing in mode 2 means that HTDSM even manages to overcompensate.



(a) (b)

Figure 4. Laplace DSM with ALD. The setup and figures are identical to Figure 2, except that Laplace noise is used ($\beta = 1$ in the general formulation). **a** depicts the diamond, rather than circular, noise structure of a diagonal bivariate Laplace distribution. **b** demonstrates that ALD sampling converges even with Laplace (sub-Gaussian, piece-wise differentiable) diffusion, confirming Theorem 2.1. Full details in Figure 13.



(a) (b)

Figure 5. Laplace DSM with ALD. The setup and figures are identical to Figure 14, except that Laplace noise is used. **b** demonstrates that Laplace noise compensates for the class imbalance. 29.5% of particles finishing in mode 2 even manages to overcompensate.

To confirm that this trend is present for all $\beta < 2$ noise types, Figure 6 extends this, repeating the experiment to estimate a confidence interval. Overall, the large jumps in sampling (similar to Lévy flights) free the sampling paths from being dominated by the more populous mode while preserving useful score estimates which point toward the underrepresented local maximum of the PDF.

Another insight is offered in Table 1, where DSM and HTDSM are used with Gaussian or Laplace diffusion. As expected, models trained with standard DSM diverge when Laplace diffusion is used for sampling[2]. Also consistent with Figures 14 and 6, DSM with Gaussian ALD suffers from mode collapse. Nevertheless, HTDSM can use sub-Gaussian diffusion to overcompensate for the asymmetric data, a trait that is valuable for realistic scenarios which often contain class imbalances. Finally, HTDSM can be

[2]Diverging here refers to approaching very large values which completely ignore the distribution modes (even if they are technically closer to one of the two).

used with Gaussian ALD solely as a method for providing better score estimates. This final process leads to the most accurate estimate of the imbalance in Table 1 and suggests that the way forward is to leverage the stability of Gaussian ALD alongside HTDSM gradients which are likely to be more accurate in low probability regions.

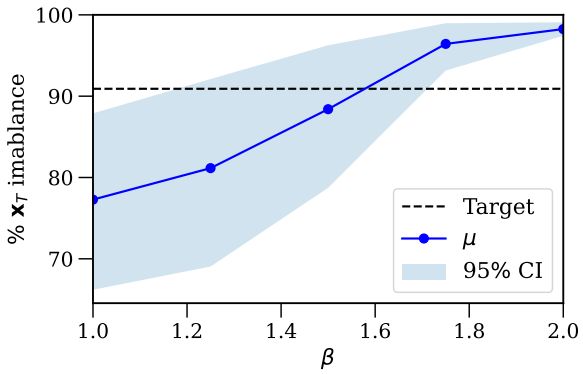


Figure 6. Mean percentage imbalance of generated data for $\beta \in [1, 2]$ across 10 runs of the 10:1 experiment in Figure 14. Noise in training and sampling is sub-Gaussian. The 95% confidence interval is bootstrapped from 10,000 resamples.

	Gaussian diffusion	Laplace diffusion
DSM	98.25 (97.40, 99.09)	Divergent
HTDSM	**87.44 (82.85, 91.50)**	77.28 (66.38, 87.79)

Table 1. Mean (plus lower/upper bounds for a 95% confidence interval bootstrapped from 10,000 resamples) percentage imbalance of generated data for the 10:1 experiment in Figure 14 with Gaussian or Laplace DSM or diffusion. DSM with Laplace noise at sampling time diverges regularly due to inaccurate gradients.

The central take-aways of these synthetic experiments are:

- HTDSM is an efficacious method of estimating a distribution's score function.
- Sub-Gaussian diffusion, causing Lévy-flight-like sampling paths, can overcome class imbalances, motivating extension to the continuous case (see Section A.3). Sufficiently accurate and compensatory score estimates for these paths can also only be achieved with HTDSM.
- HTDSM with Gaussian diffusion offers a potentially even more general solution.

3.2. High-dimensional class imbalances

When training each SBM, multiple noise levels were used to allow for ALD at sampling time. Each noise level adds generalised normal distribution noise to the image, scaled by the constant factored into the derivations in Section 2. For an input image and its sampled noise, the generalised normal score is calculated according to (15), and set as the model target. The time complexity impact of sampling $\mathcal{GN}(\mu = x, \alpha = 1, \beta)$ noise at scale is minimal, as the sampling procedure for each dimension is simply

$$\gamma \sim \text{Gamma}\left(\text{shape} = 1 + 1/\beta, \text{rate} = 2^{-\beta/2}\right) \tag{32}$$

$$\delta = \alpha\gamma^{1/\beta}/\sqrt{2} \tag{33}$$

$$\hat{x} \sim \mathcal{U}(\mu - \delta, \mu + \delta), \tag{34}$$

following Choy & Walker (2003), where $\mathcal{U}$ denotes the uniform distribution.

At sampling time, in-line with Song & Ermon (2019), images are initialised by a uniform distribution over the pixels, before ALD iteration begins. ALD (see Algorithm 2), uses the multiple noise levels $\sigma_1, \ldots, \sigma_k$ from training time, with a learning rate proportional to the ratio of the squared current noised level to the squared maximum noise level. Each noise level iterates for step limits ranging from 10 to 500, depending on both the dataset and the value of β—the latter due to the low absolute score values for distant noise when $\beta < 1.5$ (see Figure 3(b)).

3.2.1. MNIST 1 vs. 8.

To extend analysis of how HTDSM mitigates class imbalances in higher dimensions, Figure 7 and 8 present model generation results for a simplified version of the MNIST dataset. The data is limited to contain only the classes 1 and 8, which were chosen for their contrast in pixel space. The goal of Figure 7 was to demonstrate how Gaussian DSM SBMs perform poorly with ALD in the presence of asymmetric class representation, by inducing an imbalance between classes 1 and 8. However, Figure 7(a) demonstrates mode collapse before the class ratio is even manipulated, and is also supported by the more expected imbalance in Figure 7(b). Gaussian DSM suffering such issues in this minimal setting appears to contradict Song & Ermon (2019), where the motivation for combining DL and ALD was to overcome uneven mode weights. Moreover, it brings into question the cause of recent impressive generative results with SBMs, which may require the regularisation of many classes in the data to produce more general score estimates.

To reinforce this result, the same even-class model was reran to produce 100 samples[3]. DSM with 100 steps per level (s/l) produced six ones with $P(6) < 10^{-21}$ under a binomial model, whereas HTDSM produced eighteen ones with $P(18) < 10^{-10}$, a massive relative improvement. As the generated HTDSM images were speckled, the same experiment was repeated with 1,000s/l. This revealed that more

[3]Models were also retrained to verify that this issue was reproducible.

sampling steps alleviates Gaussian DSM imbalance almost completely, producing 48 ones, and for HTDSM 32 ones were generated (well within two standard deviations of the normal approximation to the underlying binomial distribution here). Therefore, one conclusion is that HTDSM is beneficial for varied sampling in compute-constrained scenarios and that avoidance of mode collapse in the literature may be, in part, due to intensive sampling procedures at high noise levels.

Figure 8 demonstrates these heavier sampling results for HTDSM. Generated digits for a HTDSM model are trained in-line with Figure 7 and sampled with varied diffusion type and steps per level (s/l). In Figure 8(a), speckle is observed, whereas more sampling steps in Figure 8(b) leads to a more even class balance. The difference between Figures 8(c) and 8(d) confirms that sub-Gaussian diffusion can be used in high dimensions successfully, as long as the number of sampling steps is increased.

(a) Class ratio 1:1.

(b) Class ratio 2:1.

Figure 7. Generated digits for a Gaussian DSM model trained for 20,000 steps on digits 1 and 8 from the MNIST dataset and sampled with 100 steps per level (s/l) of Gaussian ALD. No digits resembling a 1 are present for 25 samples, indicating a sampling process which induces (**a**) or exacerbates (**b**) the class imbalance.

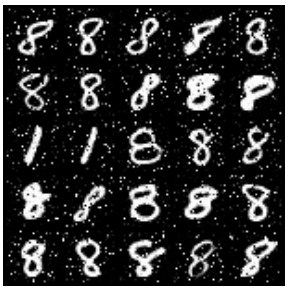
3.3. High-dimensional unconditional generation

Tables 2 and 3 summarise the DGM metrics attained using HTDSM on the MNIST and Fashion-MNIST datasets. For metrics reliant upon Inception v3 (IS, FID, and KID, see Section B.3), 10,000 samples were generated and compared to the respective training dataset using the Python `torch-fidelity`[4] package. Also, precision, recall, density, and coverage were estimated (with the same samples) using the Python `prdc`[5] package with the number of nearest neighbours, k, set to 5 (Naeem et al., 2020).

For HTDSM on MNIST with $\beta = 1.5$, precision, recall, and coverage were found to improve over standard DSM. However, all other metrics did not improve and Inception-based metrics are markedly worse. Although this suggests lower perceptual quality, Figure 9 seems to refute this. In particular, Figure 9(b) depicts Gaussian ALD failing to generate a single digit similar to a one, and the probability of the seventeen zeros occurring in the real dataset is less than 10^{-7}, so problems persist. In Table 3, $\beta = 1.5$ dominates the majority of the metrics, demonstrating that HTDSM is advantageous for certain datasets. Overall, $\beta < 2$ is a promising direction for future research and larger-scale experiments, but $\beta = 1.0$ suffered from convergence issues at scale. It is also possible to explore the effects of light-tailed DSM by setting $\beta = 2.5$. As this corresponds to estimating a score function which is very large for high noise (see Figure 3(b) for intuition), diffusion convergence was often found to be too quick, resulting in cartoon-like final images with strong features and no subtleties.

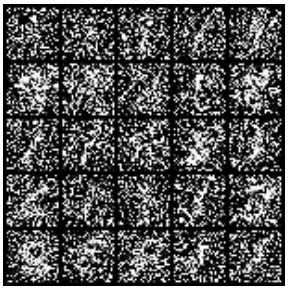
[4] https://github.com/toshas/torch-fidelity

[5] https://github.com/clovaai/generative-evaluation-prdc



(a) ALD, 100 s/l.

(b) ALD, 1,000 s/l.



(c) Laplace ALD, 100 s/l.

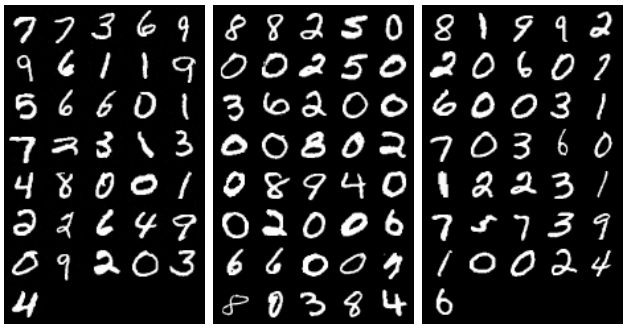
(d) Laplace ALD, 1,000 s/l.

Figure 8. Generated digits for a HTDSM model trained in-line with Figure 7 and sampled with varied diffusion type and steps per level (s/l). In **a**, speckle is observed, whereas more sampling steps in **b** leads to a more even class balance. The difference between **c** and **d** confirms that sub-Gaussian diffusion can be used in high dimensions successfully, as long as the number of sampling steps is increased.

4. Summary and outlook

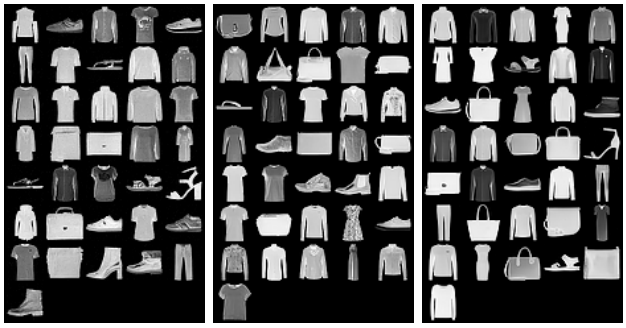
4.1. Summary

This paper has provided a thorough expansion of the theory behind discrete-level denoising score matching for score-based models. An in-depth discussion of SBM research was



(a) $\beta = 1.5$. (b) $\beta = 2.0$. (c) $\beta = 2.5$

Figure 9. Unconditional samples from a model trained with HTDSM, and sampled from using Gaussian ALD, for different values of β on the MNIST dataset.



(a) $\beta = 1.5$. (b) $\beta = 2.0$. (c) $\beta = 2.5$

Figure 10. Unconditional samples from a model trained with HTDSM, and sampled from using Gaussian ALD, for different values of β on the Fashion-MNIST dataset.

provided in Section 1, leading up to very recent progress. This foothold was then used as the basis for novel theoretical expansion of learning with DSM to heavy-tailed DSM, noising and denoising with the family of generalised normal distributions. Insight into the undesirable n-dimensional annuli in Gaussian DSM, as well as an understanding of the generalised normal score function for $\beta < 2$, motivated the use of heavier tails. GN noise was then found to concentrate in a skewed distribution which prompted a general algorithm to choose a noise scaling sequence in Section A.2.

Examples in low and high dimensions demonstrated the propensity of ALD with Gaussian DSM to suffer from mode collapse—a phenomenon easily exacerbated by class imbalances. The thorough $2D$ example, explored in Section 3.1, outlined differences at both training and sampling time for DSM and HTDSM. The latter was shown to be a tenable alternative method of estimating a distribution's score function. Experiments suggested that heavy-tailed noise can always be scaled down to dampen sampling with jumps, whereas limitations such as class imbalances are inherent to the data. HTDSM with Gaussian diffusion seems to offer the most general method of learning and sampling, balancing better gradients in low probability regions with well-behaved diffusion.

Despite stability issues in the underlying implementation, when scaling to higher dimension datasets, HTDSM continued to offer promising results. In particular, $1 < \beta < 2$ appears to be a relatively stable type of GN noise which potentially offers improved image generation according to a range of metrics in Section 3.3. Moreover, sub-Gaussian diffusion, causing Lévy-flight-like sampling paths, can also overcome class imbalances, motivating extension to the continuous case (see Section A.3), while sufficiently accurate and compensatory score estimates for these paths can only be achieved with HTDSM.

4.2. Outlook

Of particular note for future research is the shape of the high-dimensional heavy-tailed noise in Figure 4(a). As mentioned in Section 3.1, the use of a radial basis for the noise distribution, similar to that used in Farquhar et al. (2020) for Bayesian neural network parameterisation, offers a more natural approach to noising. Alternatively, in the case of highly axis-aligned data, it is possible to consider heavy tails extending along the basis axes. This approach has further merit after projecting data onto its most influential principal components, or could be used after mapping the data to any more natural underlying geometry with axis-alignment.

Once the stability issues of Section 3.3 are resolved, several practical steps can also be taken beyond the empirical evidence of this work. The more general method for initialising noise levels from Section A.2 can be explored in practice. Secondly, currently inconclusive results on CIFAR10 can be expanded to properly compare Inception distances across training methods because their use on MNIST and Fashion-MNIST is relatively unusual and debatable (although extant in the literature). Finally, the experiments of Section 3 can be scaled up to much larger image datasets such as CelebA, FFHQ, and beyond.

As discussed in Section A.3, the continuous extension of HTDSM has several expected properties which remain an active area of research following this work. The main theoretical goal is to describe how sub-Gaussian diffusion can be reversed by another (potentially different) diffusion. To do this, it is necessary to investigate the general, non-Brownian, form of the Kolmogorov backward equations, in an analysis beyond that of Song et al. (2020b) and this paper. As well as the suggestions made in Section A.3, it is also of note that this line of work has rich potential links with fractional Brownian motion (Lévy, 1953) and the fractional Fokker-Planck equation (Metzler et al., 1999).

	$\beta = 1.0$	$\beta = 1.5$	$\beta = 2.0$	$\beta = 2.5$
Precision ↑	0.0	**0.919**	0.912	0.898
Recall ↑	1.0	0.929	**0.936**	0.936
Density ↑	0.0	**0.906**	0.867	0.868
Coverage ↑	0.0	**0.892**	0.780	0.661
IS ↑	1.303 ± 0.006	1.942 ± 0.017	$\mathbf{2.037 \pm 0.037}$	1.922 ± 0.030
KID ↓	0.569 ± 0.003	0.075 ± 0.002	$\mathbf{0.016 \pm 0.002}$	0.032 ± 0.002
FID ↓	375.251	54.611	**19.399**	31.847

Table 2. DGM metrics for unconditional samples from a model trained with HTDSM, and sampled from using Gaussian ALD, for different values of β on the MNIST dataset. Arrows indicate whether higher (↑) or lower (↓) metric values are better.

	$\beta = \mathbf{1.0}$	$\beta = 1.5$	$\beta = 2.0$	$\beta = 2.5$
Precision ↑	0.0	0.905	0.925	**0.953**
Recall ↑	1.0	**0.755**	0.747	0.721
Density ↑	0.0	1.323	1.551	**1.878**
Coverage ↑	0.0	**0.840**	0.692	0.567
IS ↑	1.303 ± 0.006	$\mathbf{4.108 \pm 0.102}$	3.585 ± 0.085	3.170 ± 0.074
KID ↓	0.569 ± 0.003	$\mathbf{0.022 \pm 0.001}$	0.026 ± 0.001	0.042 ± 0.002
FID ↓	375.251	**32.990**	41.661	63.039

Table 3. DGM metrics for unconditional samples from a model trained with HTDSM, and sampled from using Gaussian ALD, for different values of β on the Fashion-MNIST dataset.

References

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Barron, A. R. Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133, 1994.

Besag, J. Comments on "representations of knowledge in complex systems" by u. grenander and mi miller. *J. Roy. Statist. Soc. Ser. B*, 56:591–592, 1994.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Bishop, C. M. et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

Box, G. E. and Tiao, G. C. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

Cheng, X., Chatterji, N. S., Abbasi-Yadkori, Y., Bartlett, P. L., and Jordan, M. I. Sharp convergence rates for langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

Choy, S. B. and Walker, S. G. The extended exponential power distribution and bayesian robustness. *Statistics & probability letters*, 65(3):227–232, 2003.

Dytso, A., Bustin, R., Poor, H. V., and Shamai, S. Analytical properties of generalized gaussian distributions. *Journal of Statistical Distributions and Applications*, 5(1):1–40, 2018.

Eicker, F. Sums of independent squared cauchy variables grow quadratically: Applications. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 133–140, 1985.

Farquhar, S., Osborne, M. A., and Gal, Y. Radial bayesian neural networks: beyond discrete support in large-scale bayesian deep learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1352–1362. PMLR, 2020.

Gil, A., Segura, J., and Temme, N. M. Efficient and accurate algorithms for the computation and inversion of the incomplete gamma function ratios. *SIAM Journal on Scientific Computing*, 34(6):A2965–A2981, 2012.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., and Welling, M. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Huang, C.-W., Lim, J. H., and Courville, A. A variational perspective on diffusion-based generative models and score matching. *arXiv preprint arXiv:2106.02808*, 2021.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Jolicoeur-Martineau, A., Li, K., Piché-Taillefer, R., Kachman, T., and Mitliagkas, I. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Kim, D., Shin, S., Song, K., Kang, W., and Moon, I.-C. Score matching model for unbounded data score. *arXiv preprint arXiv:2106.05527*, 2021.

Kingma, D. P., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. *arXiv preprint arXiv:2107.00630*, 2021.

Kobyzev, I., Prince, S., and Brubaker, M. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Lévy, P. *Random functions: general theory with special reference to Laplacian random functions*, volume 1. University of California Press, 1953.

Mandelbrot, B. B. and Mandelbrot, B. B. *The fractal geometry of nature*, volume 1. WH freeman New York, 1982.

Meng, C., Song, Y., Li, W., and Ermon, S. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems*, 34, 2021.

Metzler, R., Barkai, E., and Klafter, J. Anomalous diffusion and relaxation close to thermal equilibrium: A fractional fokker-planck equation approach. *Physical review letters*, 82(18):3563, 1999.

Mou, W., Ma, Y.-A., Wainwright, M. J., Bartlett, P. L., and Jordan, M. I. High-order langevin diffusion yields an accelerated mcmc algorithm. *arXiv preprint arXiv:1908.10859*, 2019.

Nachmani, E., Roman, R. S., and Wolf, L. Non gaussian denoising diffusion models. *arXiv preprint arXiv:2106.07582*, 2021.

Nadarajah, S. A generalized normal distribution. *Journal of Applied statistics*, 32(7):685–694, 2005.

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pp. 7176–7185. PMLR, 2020.

Neal, R. M. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Øksendal, B. Stochastic differential equations. In *Stochastic differential equations*, pp. 65–84. Springer, 2003.

Papamakarios, G., Pavlakou, T., and Murray, I. Masked autoregressive flow for density estimation. *arXiv preprint arXiv:1705.07057*, 2017.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. Assessing generative models via precision and recall. *arXiv preprint arXiv:1806.00035*, 2018.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.

Särkkä, S. and Solin, A. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

Şimşekli, U. Fractional langevin monte carlo: Exploring lévy driven stochastic differential equations for markov chain monte carlo. In *International Conference on Machine Learning*, pp. 3200–3209. PMLR, 2017.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *arXiv preprint arXiv:2006.09011*, 2020.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pp. 574–584. PMLR, 2020a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Stacy, E. W. A generalization of the gamma distribution. *The Annals of mathematical statistics*, pp. 1187–1192, 1962.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Talagrand, M. A new look at independence. *The Annals of probability*, pp. 1–34, 1996.

Vahdat, A., Kreis, K., and Kautz, J. Score-based generative modeling in latent space. *arXiv preprint arXiv:2106.05931*, 2021.

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

A. Derivations

A.1. Unintuitive high-dimensional statistics

To demonstrate unintuitive statistical behaviour in high-dimensional space, consider two classic examples:

- For an iid random vector $\mathbf{X} = [X_1, \ldots, X_n]^{\mathrm{T}}$ with $X_i \sim U(a, b)$, the majority of probability mass resides in the corners of the hypercube for high n.
- For an iid random vector $\mathbf{X} = [X_1, \ldots, X_n]^{\mathrm{T}}$ with $X_i \sim \mathcal{N}(\mu, \sigma)$, the probability mass concentrates in a thin annulus (shell).

Both of these phenomena are instances of the *concentration of measure*—the principle that a random variable that depends in a Lipschitz way on many independent variables is essentially constant (Talagrand, 1996). As the latter spherical case motivates this work, it is now explored in full.

Consider the *squared L^2 length distribution* of the isotropic Gaussian vector

$$Y = \|\mathbf{X}\|_2^2, \tag{35}$$

which, by independence, gives

$$Y = \sum_{i=1}^{n} X_i^2 = nX_i^2. \tag{36}$$

Changing random variables

$$\begin{aligned} F_Y(y) &= P(Y \leq y) && (37)\\ &= P(nX_i^2 \leq y) && (38)\\ &= P\left(|X_i| \leq \sqrt{\frac{y}{n}}\right) && (39)\\ &= \Phi\left(\sqrt{\frac{y}{n}}\right) - \Phi\left(-\sqrt{\frac{y}{n}}\right), && (40) \end{aligned}$$

where Φ is the Gaussian CDF, and differentiating gives the PDF of Y

$$\begin{aligned} f_Y(y) = F_Y'(y) &= \frac{1}{2\sqrt{ny}}\phi\left(\sqrt{\frac{y}{n}}\right) - \frac{1}{2\sqrt{ny}}\phi\left(-\sqrt{\frac{y}{n}}\right) && (41)\\ &= \frac{1}{\sqrt{ny}}\phi\left(\sqrt{\frac{y}{n}}\right) && (42)\\ &= \frac{1}{\sqrt{ny}}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{y}{2n}\right\}, && (43) \end{aligned}$$

where ϕ is the Gaussian PDF. Rewriting this expression and using $\sqrt{\pi} = \Gamma(1/2)$

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{\pi}\sqrt{2n}}y^{-\frac{1}{2}}\exp\left\{-\frac{y}{2n}\right\} && (44)\\ &= \frac{1}{\Gamma(1/2)(2n)^{\frac{1}{2}}}y^{\frac{1}{2}-1}\exp\left\{-\frac{y}{2n}\right\}, && (45) \end{aligned}$$

recovers the gamma distribution PDF

$$W \sim \text{Gamma}(k, \theta) \implies f(w; k, \theta) = \frac{1}{\Gamma(k)\theta^k}w^{k-1}e^{-\frac{k}{\theta}}, \tag{46}$$

and demonstrates that Y is distributed as either of

$$Y \sim \text{Gamma}\left(k = \frac{1}{2}, \theta = 2n\right) = n\text{Gamma}\left(\frac{1}{2}, 2\right), \tag{47}$$

and therefore follows either of the chi-squared distributions

$$Y \sim \chi^2(n) = n\chi^2(1), \tag{48}$$

which approaches a Gaussian distribution centred at n in the limit $n \to \infty$. See Section 2.1.4 for a full description of the moments and Figure 15 for a visualisation of the chi-squared distribution for increasing degrees of freedom.

A.2. Scale parameter sequences for arbitrary noise distributions

It is now apparent that the generalised normal distribution can be used for denoising score matching, leading to thicker concentric annuli. However, the non-zero skew in (26), representing the asymmetric length distribution for low n, cannot necessarily be ignored. This asymmetry implies that the spacing of noise levels using variance in (Song & Ermon, 2020), is inaccurate.

In particular, the probability mass in the left/right tail of one noise level annulus will be larger than the probability mass in the right/left tail, respectively, of the adjacent annulus. In practical terms, this means overly-dense concentric noise levels in ALD. Although the time needed to sample each noise-level per training iteration will not be affected, this will increase overall training time, sampling (generation) time, and render some sampling steps redundant. As the same problem extends to the generalised noise characterised in this paper, it is vital to examine the skew of the length scale distribution from (23).

To begin, note the rth raw moment of $Y \sim \mathcal{GG}(a, d, p)$ is

$$\mathbb{E}\left[Y^r\right] = a^r \frac{\Gamma((d+r)/p)}{\Gamma(d/p)}, \tag{49}$$

implying that, for $\|\mathbf{X}\|_2^2 = Y \sim \ \mathcal{GG}(a = n, d = 1/2, p = \beta/2)$

$$\mathbb{E}\left[Y^3\right] = n^3 \frac{\Gamma(7/\beta)}{\Gamma(1/\beta)}. \tag{50}$$

Then, using the 3^{rd} central moment expansion for skew

$$\begin{aligned}
\text{Skew}(Y) &= \mathbb{E}\left[\left(\frac{Y-\mu}{\sigma}\right)^3\right] && (51)\\
&= \frac{1}{\sigma^3}\left(\mathbb{E}\left[Y^3\right] - 3\mu\mathbb{E}\left[Y^2\right] + 3\mu^2\mathbb{E}[Y] - \mu^3\right) && (52)\\
&= \frac{1}{\sigma^3}\left(\mathbb{E}\left[Y^3\right] - 3\mu\sigma^2 - \mu^3\right) && (53)\\
&= \frac{1}{n^3 C_2^{\frac{3}{2}}}\left\{n^3\frac{\Gamma(7/\beta)}{\Gamma(1/\beta)} - 3nC_1n^2C_2 - n^3C_1^3\right\} && (54)\\
&= \frac{1}{C_2^{\frac{3}{2}}}\left\{\frac{\Gamma(7/\beta)}{\Gamma(1/\beta)} - 3C_1C_2 - C_1^3\right\}, && (55)
\end{aligned}$$

where C_1 and C_2 are defined as in (30) and (31) respectively. It is interesting to note that this skew expression is constant with respect to dimension.

The tangible Laplace noising, $\beta = 1$, example can now be exemplified, as in this case

$$\begin{aligned}
C_1 &= \frac{\Gamma(3)}{\Gamma(1)} = 2 && (56)\\
C_2 &= \frac{\Gamma(5)}{\Gamma(1)} - \left(\frac{\Gamma(3)}{\Gamma(1)}\right)^2 = 20, && (57)
\end{aligned}$$

and it is clear that

$$\text{Skew}(Y;\beta=1) = 20^{-3/2}\left(720 - 3\times 2\times 20 - 2^3\right) \tag{58}$$

$$= 74/\sqrt{5}^3 \tag{59}$$

$$\approx 6.19, \tag{60}$$

so the resulting distribution is very positively skewed.

Despite the disappointing prospects of this result, the potential for large asymmetric annulus overlap, it also motivates a better understanding of the general case. How can concentric annuli be constructed with equal overlapping probability mass?

To motivate the general algorithm, assess the case where the generalised normal noise is scaled by an arbitrary noise level σ_i.

$$X_i/\sigma_i \sim \mathcal{GN}(\mu=0,\alpha=1,\beta) \implies X_i \sim \mathcal{GN}(0,\sigma_i,\beta), \tag{61}$$

it is, therefore, true that for GN noise vector $\mathbf{X}$

$$\|\mathbf{X}\|_2^2 = Y \sim \mathcal{GG}\left(n\sigma_i^2, 1/2, \beta/2\right). \tag{62}$$

In an ascending sequence of noise where the goal is to calculate σ_{i+1} from σ_i with a given probability mass overlap, the quantile function of the length distribution must then be used by inverting the corresponding CDF. Here, the CDF is

$$F_{\mathcal{GG}}(x;a,d,p) = \frac{\gamma\left(d/p,(x/a)^p\right)}{\Gamma(d/p)}, \tag{63}$$

where $\gamma(\cdot)$ is the *lower incomplete gamma function*

$$\gamma(s,x) = \int_0^x t^{s-1}e^{-t}dt. \tag{64}$$

Although (63) appears difficult to invert, due to the *inverse of composite functions*, the quantile function for quantile q follows as

$$F^{-1}_{\mathcal{GG}}(q;a,d,p) = a\left[G^{-1}(q)\right]^{1/p}, \tag{65}$$

where

$$G(x) = F_{\text{Gamma}}\left(x;\alpha'=d/p,\beta'=1\right) = \frac{\gamma(\alpha',\beta' x)}{\Gamma(\alpha')} \tag{66}$$

$$= \frac{\gamma(d/p,x)}{\Gamma(d/p)}, \tag{67}$$

a scaled Gamma distribution CDF (Greek letters here are for the *standard* Gamma distribution), and the form $\gamma(c_1,c_2x)/\Gamma(c_1),\ c_1,c_2\in\mathbb{R}_+$, is known as the *regularised gamma function*.

Finally, substituting (67) into (65), gives

$$F^{-1}_{\mathcal{GG}}(q;a,d,p) = a\left(\left[\frac{\gamma(d/p,q)}{\Gamma(d/p)}\right]^{-1}\right)^{1/p}, \tag{68}$$

before substituting (62) as well provides

$$F^{-1}_{Y}(q) = n\sigma_i^2\left(\left[\frac{\gamma(1/\beta,q)}{\Gamma(1/\beta)}\right]^{-1}\right)^{2/\beta}. \tag{69}$$

After these steps, for an example overlap of 5% of probability mass, it is now possible to say

$$q_{i,0.95} = n\sigma_i^2\left(\left[\frac{\gamma(1/\beta,0.95)}{\Gamma(1/\beta)}\right]^{-1}\right)^{2/\beta}, \tag{70}$$

is the upper quantile for σ_i. Crucially, this expression can be inverted to obtain σ_{i+1} by setting the next lower bound equal to the current upper bound, $q_{i+1,0.05} = q_{i,0.95}$ and inverting

$$\sigma_{i+1} = \sqrt{\frac{q_{i,0.95}}{n}} \left(\left[\frac{\gamma(1/\beta, 0.05)}{\Gamma(1/\beta)} \right]^{-1} \right)^{-1/\beta}. \tag{71}$$

These last two equations outline a general procedure for consecutive noise levels with equal distribution overlap, detailed fully in Algorithm 3. Of practical significance is the Python function `scipy.special.gammaincinv` which numerically estimates the troublesome inverse regularised gamma function to arbitrary precision (Gil et al., 2012).

Algorithm 3 Scale parameter sequence generation

Input Fixed hyperparameters of piecewise log-differentiable noise distribution, non-overlapping distribution proportion $\delta \in (0, 1)$, small initial noise level $\sigma_1 > 0$, and large final noise level $\sigma_{\max}$.
Initialise $i = 1$, $q_i^l = 0$, and $q_i^u = 0$.
while $q_i^u < \sigma_{\max}$ **do**
  Calculate upper quantile $q_i^u = Q_{length}\left(\sigma_i, \frac{1+\delta}{2}\right)$ of nD length-scale distribution
  Calculate scaling needed to equate to lower quantile $q_{i+1}^l = q_i^u$
  $\sigma_{i+1} = Q_{length}^{-1}\left(\sigma_i, \frac{1-\delta}{2}\right)$
  $i = i + 1$
end while

A.3. Continuous extension to stochastic differential equations

Given the success of multiple noise scales in Gaussian ALD, recent SBM continuations have considered infinitely many noise levels, such that the perturbed data distributions evolve according to a stochastic differential equation (SDE). The goal is to construct a diffusion process $\{\mathbf{x}(t)\}_{t=0}^T$, $t \in [0, T]$, such that $\mathbf{x}(0) \sim p_0$ is the dataset of i.i.d. samples, and $\mathbf{x}(T) \sim p_T$ is the prior distribution, with a tractable form to generate samples efficiently. This diffusion process can be modelled as the solution to an Itô SDE[6]

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}, \tag{72}$$

where $\mathbf{w}$ is the standard Wiener process (Brownian motion), $\mathbf{f}(\cdot, t) : \mathbb{R}^n \to \mathbb{R}^n$ is a vector-valued function called the *drift* coefficient of $\mathbf{x}(t)$, and $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is a scalar function known as the *diffusion* coefficient of $\mathbf{x}(t)$. Here, the diffusion coefficient is assumed to be a scalar (instead of a $d \times d$ matrix) and does not depend on $\mathbf{x}$. The SDE has a unique strong solution as long as the coefficients are globally lipschitz in both state and time (Øksendal, 2003). Henceforth, the probability density of $\mathbf{x}(t)$ is denoted by $p_t(\mathbf{x})$, and $p_{st}(\mathbf{x}(t)|\mathbf{x}(s))$ denotes the transition kernel from $\mathbf{x}(s)$ to $\mathbf{x}(t)$, where $0 \leq s < t \leq T$. Typically, p_T is an unstructured prior distribution that contains no information about p_0.

It is possible to start from samples of $\mathbf{x}(T) \sim p_T$ and reverse the process to obtain samples from $\mathbf{x}(0) \sim p_0$. The main result in Anderson (1982) states that the reverse of a diffusion process is a also a diffusion process running backwards in time and given by the reverse-time SDE

$$d\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] dt + g(t)d\bar{\mathbf{w}}, \tag{73}$$

where $\bar{\mathbf{w}}$ is a reverse-time Wiener process and $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is estimated by

$$\theta^* = \arg\min_{\theta} \mathbb{E}_t \left\{ \lambda(t) \mathbb{E}_{\mathbf{x}(0)} \mathbb{E}_{\mathbf{x}(t)|\mathbf{x}(0)} \left[\| s_\theta(\mathbf{x}(t), t) - \nabla_{\mathbf{x}(t)} \log p_{0t}(\mathbf{x}(t)|\mathbf{x}(0)) \|_2^2 \right] \right\}, \tag{74}$$

for $\lambda : [0, T] \to \mathbb{R}_+$ a positive weighting function, $t \sim \mathcal{U}(0, T)$, $\mathbf{x}(0) \sim p_0(\mathbf{x})$, and $\mathbf{x}(t) \sim p_{0t}(\mathbf{x}(t)|\mathbf{x}(0))$. The overall process was given the general name score matching Langevin dynamics (SMLD) in Song et al. (2020b).

When using N noise scales, each perturbation kernel $p_{\sigma_i}(\mathbf{x}|\mathbf{x}_0)$ of SMLD can be derived from the Markov chain

$$\mathbf{x}_i = \mathbf{x}_{i-1} + \sqrt{\sigma_i^2 - \sigma_{i-1}^2}\mathbf{z}_{i-1}, \tag{75}$$

[6] A full description of the methods for calculus on stochastic processes, the foremost being Itô and Stratonovich calculus, can be found in (Särkkä & Solin, 2019).

where $i = 1, \dots, N$ and $\mathbf{z}_{i-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{x}_0 \sim p_{\text{data}}$, and $\sigma_0 = 0$ is used to simplify notation. Whereas Song et al. (2020b) proceed with Gaussian noise, the continuation with sub-Gaussian noise is now assessed.

To begin, let the elements of $\mathbf{l}_{i-1}$ follow a sub-Gaussian distribution. Then let $\mathbf{x}(i/N) = \mathbf{x}_i$, $\sigma(i/N) = \sigma_i$, and $\mathbf{l}(i/N) = \mathbf{l}_i \; \forall i$. With $\Delta t = 1/N$, it is then possible to write

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \sqrt{\sigma^2(t + \Delta t) - \sigma^2(t)}\, \mathbf{l}(t) \tag{76}$$

$$\approx \mathbf{x}(t) + \sqrt{\frac{d\left[\sigma^2(t)\right]}{dt} \Delta t}\, \mathbf{l}(t), \tag{77}$$

where the approximate equality holds when $\Delta t \ll 1$. In the limit $\Delta t \to 0$, this converges to

$$d\mathbf{x} = \sqrt{\frac{d\left[\sigma^2(t)\right]}{dt}}\, d\boldsymbol{\ell}(t), \tag{78}$$

where $\boldsymbol{\ell}(t)$ is a Lévy process, rather than the Wiener process $\mathbf{w}(t)$ of Song et al. (2020b) which can be solved in closed-form as an affine Brownian motion SDE.

The addition of $\boldsymbol{\ell}$ to the more formal version of the diffusion process in (72) gives

$$\mathbf{x}(t) = \int_0^t \mathbf{f}(\mathbf{x}, s) ds + \int_0^t g(s) d\boldsymbol{\ell}, \tag{79}$$

where the latter term can be interpreted as

$$\lim_{\Delta t \to 0} \left[\sum_i g(t_i)(\boldsymbol{\ell}(t_i + \Delta t) - \boldsymbol{l}(t_i)) \right]. \tag{80}$$

The sum formulation makes it clear that, for any infinitely divisible distribution[7] which sums to itself (e.g. Gaussian, Laplace, and the previously discarded Cauchy), the final distribution will be in the same family.

Therefore, it is expected that the solution to (78) describes a process which would diffuse to the underlying stable (infinitely divisible) distribution. In the context of SMLD, this means that the prior $p_T(\mathbf{x})$ need not be Gaussian and can be heavy tailed.

Unfortunately, to reverse the diffusion, it is necessary to investigate the general, non-Brownian, form of the Kolmogorov backward equations, in an analysis beyond that of Song et al. (2020b) and this paper. Instead, several practical remarks are made to finish the theory of this work.

Firstly, it is of note that the generalised normal distribution considered in this chapter is infinitely divisible for $\beta \in (0, 1] \cup \{2\}$ (Dytso et al., 2018). This result is interesting because the analysis in Section 2.1.4 suggests that $\beta < 1$ suffers from explosive and unwieldy length distribution moment coefficients, yet this region may be of theoretical intrigue for continuous HTDSM. Secondly, there exist several connections between Brownian motion and heavier-tailed diffusion through subordination—letting time evolve according to a stochastic process within another stochastic process. The prime examples of this are variance gamma (VG) processes, which can be written as a Brownian motion $W(t)$ with drift θt, subject to a random time change that follows a gamma process $\Gamma(t; 1, \nu)$

$$X^{VG}(t; \sigma, \nu, \theta) = \theta \Gamma(t; 1, \nu) + \sigma W(\Gamma(t; 1, \nu)), \tag{81}$$

where σ is a scale parameter and ν controls the time dilation. In particular, when $\nu = 1$, a VG process is equivalent to the continuous version of the $\beta = 1$ GN noise considered in this chapter. Future work may be able to use the backward Kolmogorov equation on the time-dilated Wiener process to form an ordinary differential equation describing the reverse evolution from a heavy-tailed prior to the data distribution.

[7] F is infinitely divisible if $\forall n \in \mathbb{N}, \; \exists \, n$ i.i.d. RVs s.t. $\sum_{i=1}^{n} X_i = S$ and S has the same distribution as F.

A.4. Proof of Theorem 2.1

Proof. Expanding (4) gives

$$\mathcal{J}_{ESMp}(\theta) = \int_{\mathbf{x}\in\mathbb{R}^n} p(\mathbf{x}) \left[\frac{1}{2} \underbrace{\|\nabla_{\mathbf{x}} \log p(\mathbf{x})\|_2^2}_{①} + \frac{1}{2} \underbrace{\|s_\theta(\mathbf{x})\|_2^2}_{②} - \underbrace{\nabla_{\mathbf{x}} \log p(\mathbf{x})^T s_\theta(\mathbf{x})}_{③} \right] d\mathbf{x}, \tag{82}$$

where ① can be ignored as it is constant, with no dependency on θ. For the second term, expand

$$② = \int_{\mathbf{x}\in\mathbb{R}^n} p(\mathbf{x}) \sum_{i=1}^{n} (s_\theta(\mathbf{x})_i)^2 d\mathbf{x}, \tag{83}$$

where $s_\theta(\mathbf{x})_i$ is the i^{th} component of the partial derivatives composing $s_\theta(\mathbf{x})$, and let

$$s_\theta(\mathbf{x})_i = \sum_j s_\theta(\mathbf{x})_{i,j}, \tag{84}$$

where j indexes a countable sequence of intervals partitioning the real line (except for points of zero measure). Also let each $s_\theta(\mathbf{x})_{i,j}$ be differentiable inside its corresponding interval and zero outside, permitting the derivation of

$$\int_{\mathbf{x}\in\mathbb{R}^n} p(\mathbf{x}) \sum_{i=1}^{n} (s_\theta(\mathbf{x})_i)^2 d\mathbf{x} = \sum_{i=1}^{n} \int_{\mathbf{x}\in\mathbb{R}^n} p(\mathbf{x}) \left(\sum_j s_\theta(\mathbf{x})_{i,j} \right)^2 d\mathbf{x} \tag{85}$$

$$= \sum_{i=1}^{n} \int_{\mathbf{x}\in\mathbb{R}^n} p(\mathbf{x}) \sum_j \left(s_\theta(\mathbf{x})_{i,j}\right)^2 d\mathbf{x} \tag{86}$$

$$= \int_{\mathbf{x}\in\mathbb{R}^n} p(\mathbf{x}) \sum_{i=1}^{n} \left(s_\theta(\mathbf{x})_i\right)^2 d\mathbf{x} \tag{87}$$

$$= \int_{\mathbf{x}\in\mathbb{R}^n} p(\mathbf{x}) \|s_\theta(\mathbf{x})\|_2^2 d\mathbf{x}, \tag{88}$$

the first term of (5), despite the differentiable almost everywhere formulation. This step of the proof simply shows that *when integrating, the square of the sum of piecewise non-zero functions is equal to the sum of their squares*.

③ remains, and the proof will be complete if a differentiable almost everywhere equivalent of Lemma 4 in (Hyvärinen, 2005) establishes a multivariate version of

$$\int p(x)(\log p)' f(x) dx = \int p(x) \frac{p'(x)}{p(x)} f(x) dx = \int p'(x) f(x) dx = \int p(x) f'(x) dx. \tag{89}$$

Proposition A.1. *For $i = 1$, without loss of generality (WLOG)*

$$\lim_{a\to\infty,\ b\to-\infty} [f(a, x_2, \ldots, x_n) g(a, x_2, \ldots, x_n) - f(b, x_2, \ldots, x_n) g(b, x_2, \ldots, x_n)]$$
$$= \int_{-\infty}^{\infty} f(x) \frac{\partial g(x)}{\partial x_1} dx_1 + \int_{-\infty}^{\infty} g(x) \frac{\partial f(x)}{\partial x_1} dx_1, \tag{90}$$

assuming that f is differentiable and g is differentiable almost everywhere.

Proof. WLOG break $g(x)$ into piecewise differentiable and non-zero functions along the first dimension, $g(x) = \sum_j g_j(x)$, defined in the interval I_j and zero elsewhere. Then

$$\frac{\partial f(x) g(x)}{\partial x_1} = f(x) \frac{\partial g(x)}{\partial x_1} + g(x) \frac{\partial f(x)}{\partial x_1} \tag{91}$$

$$= f(x) \frac{\partial}{\partial x_1} \left[\sum_j g_j(x) \right] + \sum_j g_j(x) \frac{\partial f(x)}{\partial x_1}, \tag{92}$$

where all variables except x_1 can be fixed. Then, integrating over $x_1 \in \mathbb{R}$,

$$[f(x)g(x)]_{-\infty}^{\infty} = \int_{-\infty}^{\infty} f(x) \sum_j \frac{\partial g_j(x)}{\partial x_1} dx_1 + \sum_j \int_{I_j} g_j(x) \frac{\partial f(x)}{\partial x_1} dx_1 \tag{93}$$

$$= \int_{-\infty}^{\infty} f(x) \frac{\partial g(x)}{\partial x_1} dx_1 + \int_{-\infty}^{\infty} g(x) \frac{\partial f(x)}{\partial x_1} dx_1, \tag{94}$$

where the first term arises by construction and the second arises via a telescoping sum. □

This proposition allows for an equivalent to the final step in (Hyvärinen, 2005)

$$\begin{aligned} -\int \frac{\partial p_{\mathbf{x}}(\mathbf{x})}{\partial x_1} s_\theta(\mathbf{x}) d\mathbf{x} &= -\int \left[\int \frac{\partial p_{\mathbf{x}}(\mathbf{x})}{x_1} s_\theta(\mathbf{x}) dx_1 \right] d(x_2, \ldots, x_n) \\ &= -\int \Bigg[\lim_{a \to \infty, b \to -\infty} [p_{\mathbf{x}}(a, x_2 \ \ldots, x_n) s_\theta(a, x_2 \ \ldots, x_n) \\ &\qquad - p_{\mathbf{x}}(b, x_2 \ \ldots, x_n) s_\theta(b, x_2 \ \ldots, x_n)] \\ &\qquad\qquad - \int \frac{s_\theta(\mathbf{x})}{\partial x_1} p_{\mathbf{x}}(\mathbf{x}) dx_1 \Bigg] d(x_2, \ldots, x_n). \end{aligned}$$

The choice of $i = 1$ is arbitrary and the limit is zero by assumption, therefore proving

$$-\int_{-\infty}^{\infty} p_{\mathbf{X}}(\mathbf{x}) \frac{\partial \log p_{\mathbf{X}}(\mathbf{x})}{\partial x_i} s_\theta(\mathbf{x})_i dx_i = \int \frac{\partial s_\theta(\mathbf{x})_i}{\partial x_i} p_{\mathbf{x}}(\mathbf{x}) dx_i, \tag{95}$$

returns the i^{th} component ③, which is summed to form the trace. □

B. Extended results

B.1. Extended $2D$ example

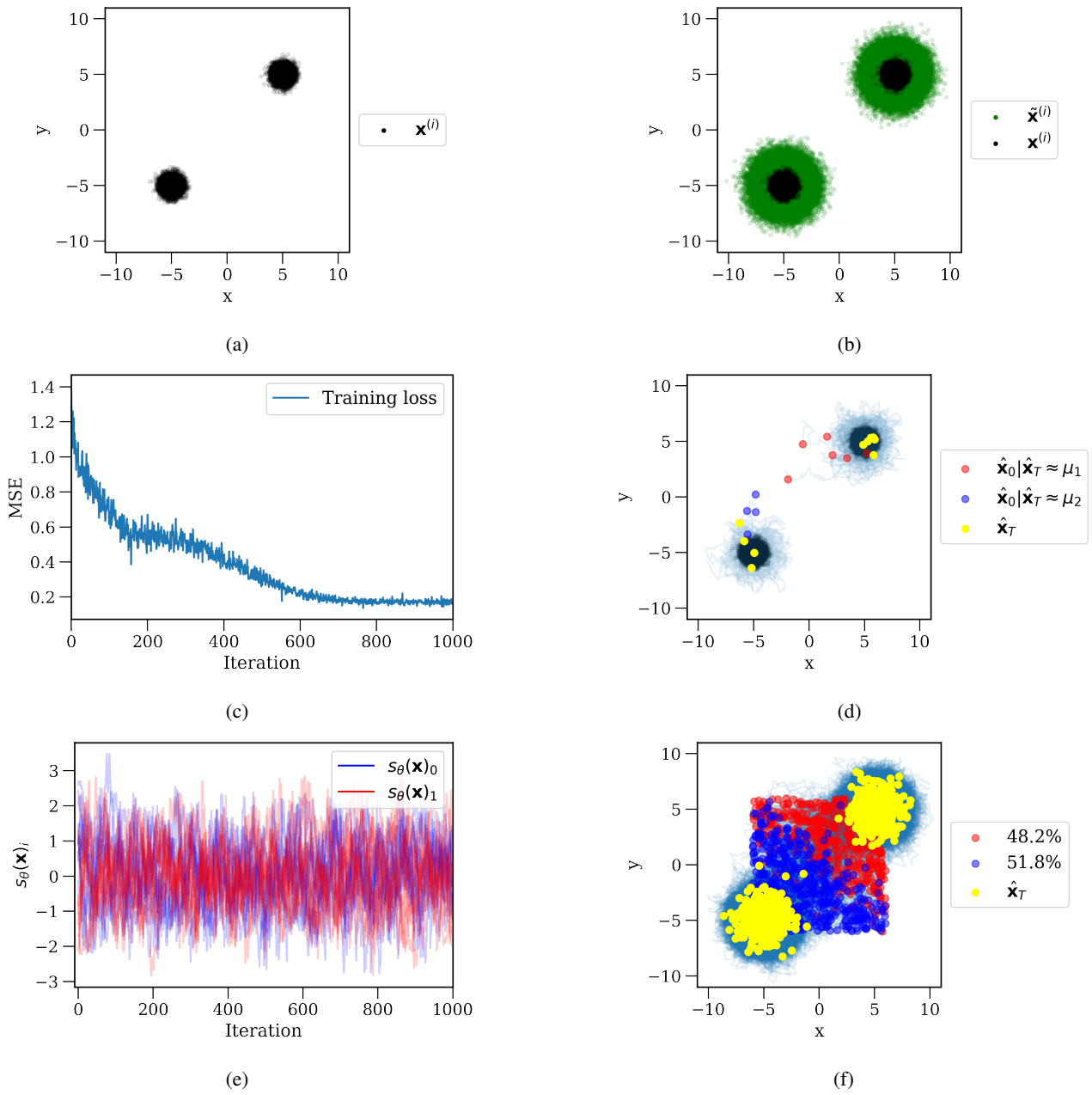


Figure 11. DSM training and LD sampling. In **a**, $p(\mathbf{x})$ is modelled as an additive mixture of ($k = 2$) bivariate Gaussians with 10,000 samples per mode. A depth 3 MLP ($2 \to 16 \to 16 \to 2$, intermediate activations ReLU, batch size 256) is trained to estimate the score from samples noised by $q_\sigma(\tilde{\mathbf{x}}|\mathbf{x}) \sim \mathcal{N}(\mathbf{x}, \mathbf{I})$. All training noise samples are shown as green in **b** and the training convergence is depicted in **c**. Then, in **d**, starting from $\hat{\mathbf{x}}_0 \sim \mathcal{U}(-6, 6) \times \mathcal{U}(-6, 6)$, 10 sampled particles are evolved to convergence using 1,000 steps of Langevin Dynamics with step size 0.1 and matching noise scale. The score estimates used during sampling are presented in **e**. Finally, the same sampling is repeated in **f** for 1,000 particles to demonstrate the decision boundary, its asymmetry (relevant for class imbalance), and the upper bound on approximation accuracy due to the underlying unit noise.

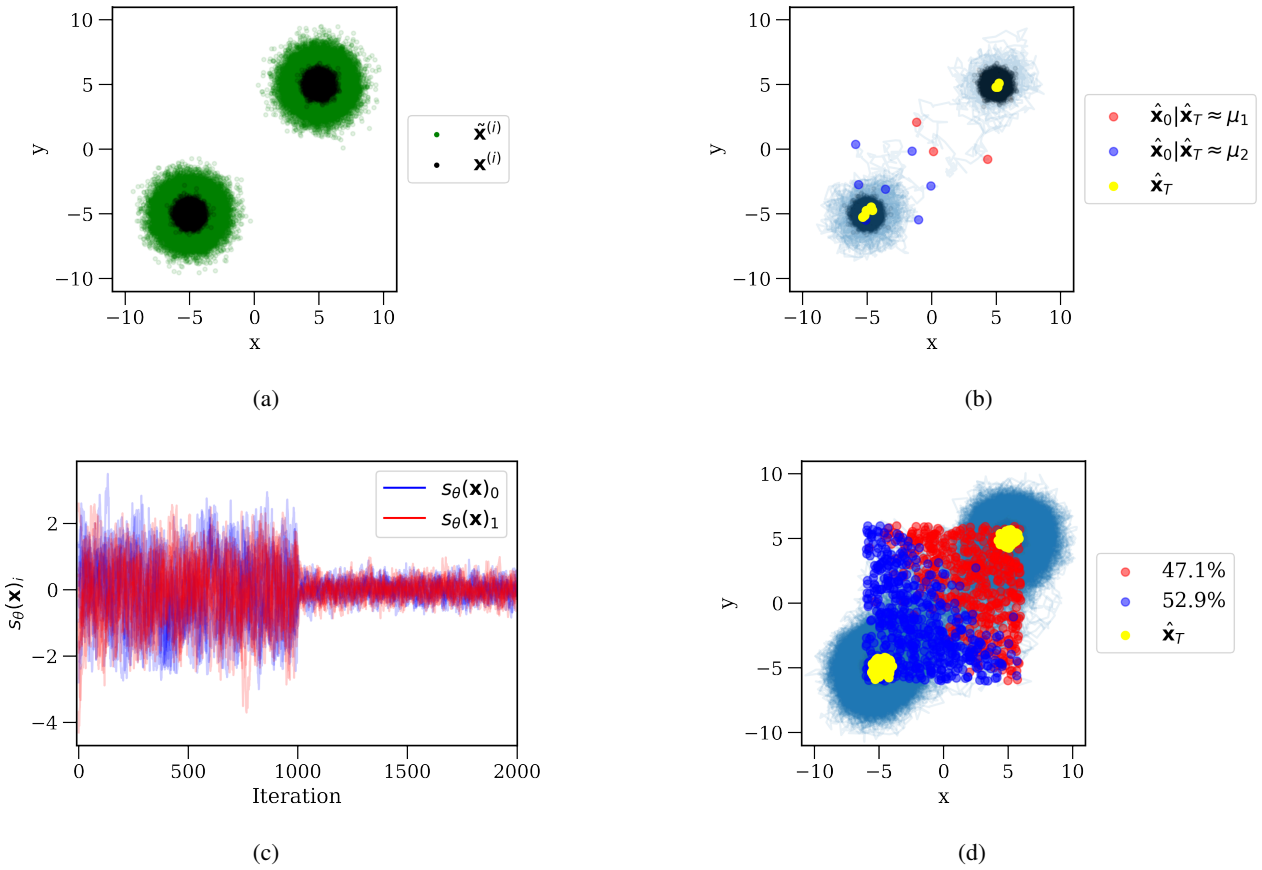


Figure 12. Multiple noise level DSM training and ALD sampling. The setup and figures are identical to Figure 11 except that two noise scales, $\sigma_1 = 1.0$ and $\sigma_2 = 0.25$, are used in training and sampling. The sequential use of decreasing noise levels in sampling can be seen in **c**. It is evident that ALD drastically improves the final distribution estimate due to the decrease in score estimate scale. It is also relevant to subsequent class imbalance problems that the sampling procedure is slightly asymmetric. For all models trained, class asymmetry is consistent across sampling runs, but not across DSM retraining, so is an artefact of the model.

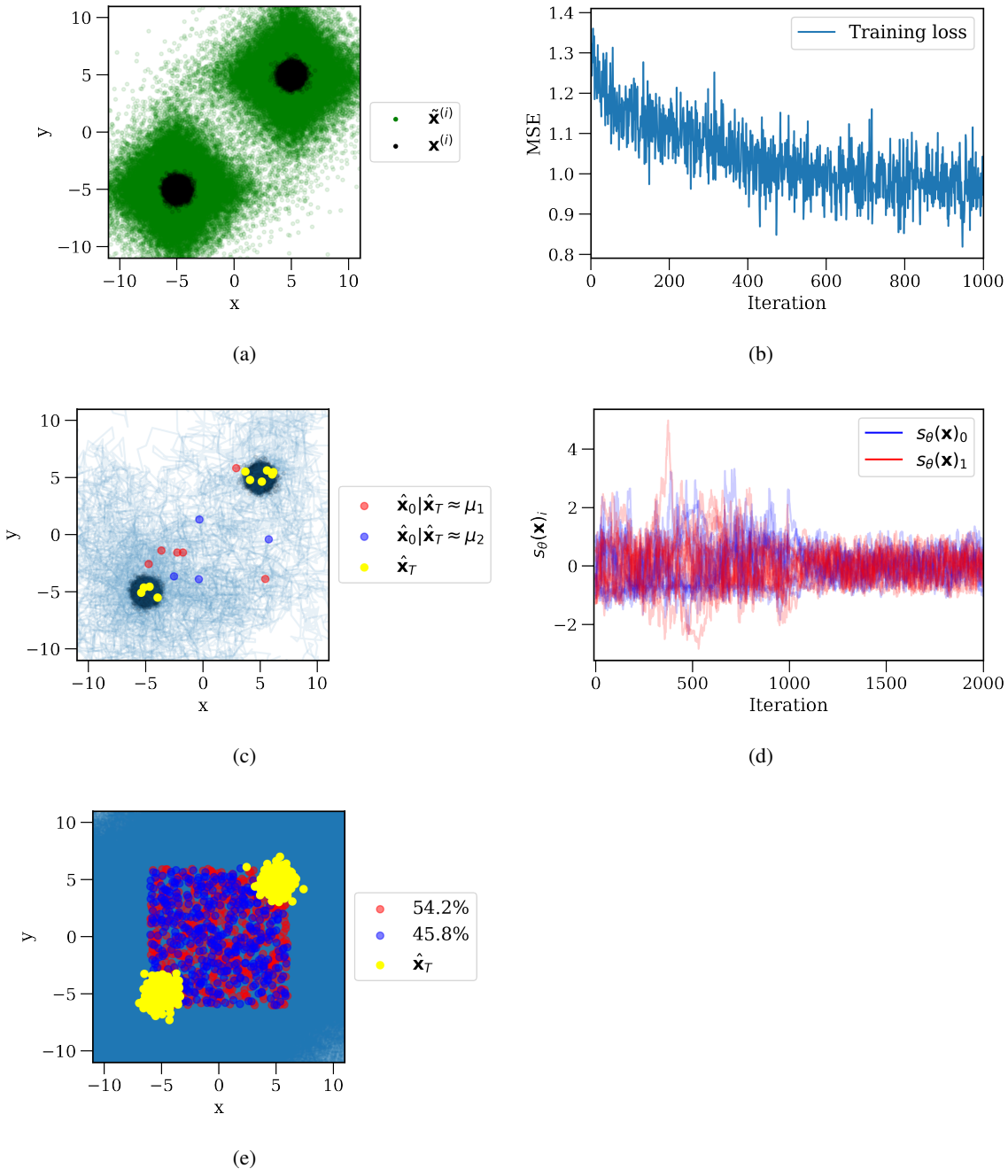


Figure 13. Laplace DSM with ALD. The setup and figures are identical to Figure 12, except that Laplace noise is used ($\beta = 1$ in the general formulation). **a** depicts the diamond, rather than circular, noise structure of a diagonal bivariate Laplace distribution. **b** and **c** respectively demonstrate that ALD training and sampling converge even with Laplace (sub-Gaussian, piece-wise differentiable) diffusion, confirming Theorem 2.1.

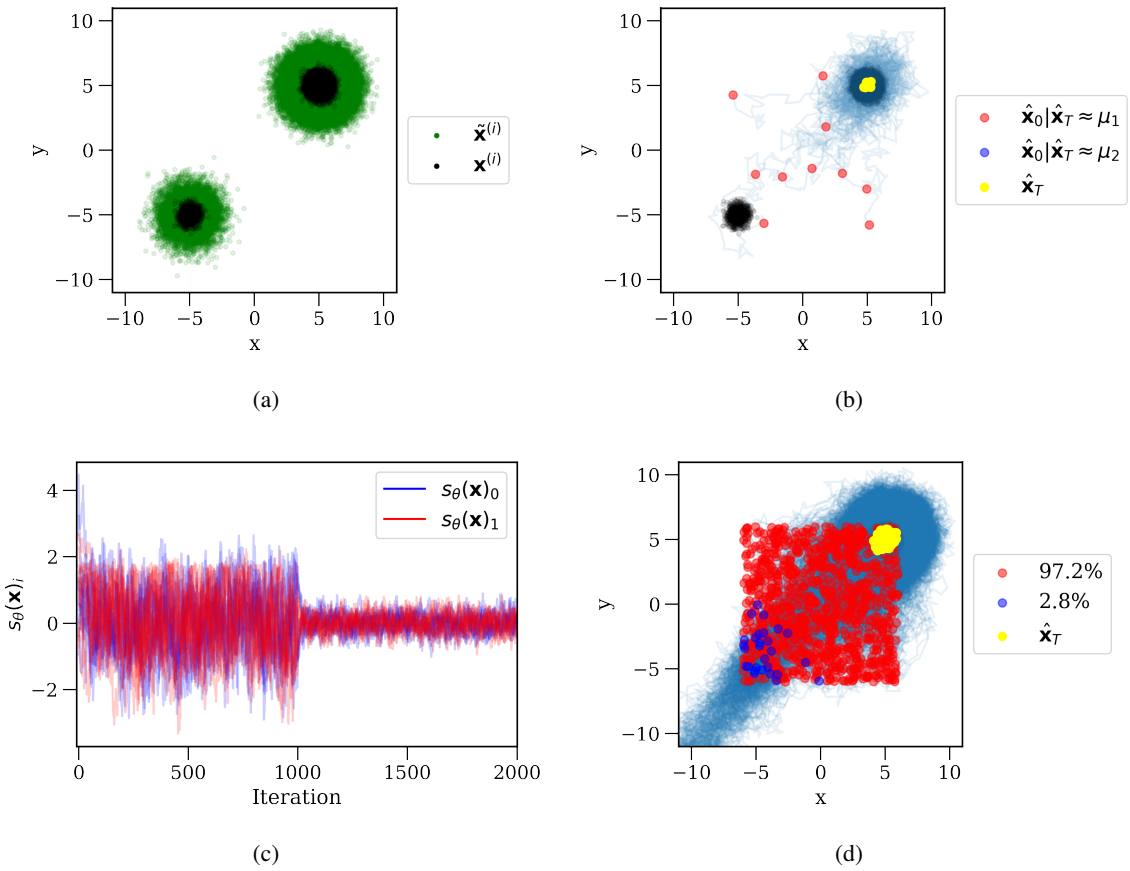


Figure 14. DSM with ALD mode collapse. The setup and figures are identical to Figure 12, except that $p(\mathbf{x})$ samples now have an imbalance of 10:1 between modes 1 (upper right) and 2 (lower left) respectively.

B.2. Experimental setup

For benchmark image datasets, the experimental setup is described for reproducibility. Different neural network architectures were used for different datasets, with respective sizes positively correlated. For MNIST, Fashion-MNIST, and CIFAR10 ($28 \times 28 \times 1$ or $32 \times 32 \times 3$), the ResNet architecture (He et al., 2016) from Song & Ermon (2020) is used in-line with the literature. In the case of CIFAR10, one convolutional layer is added during spatial down and up-sampling to reflect the more complex data. Then, for any larger datasets, experiments also follow in-line with the literature, but are limited by lower levels of compute. NN architecture specifics and further training information are summarised in Appendix B of Song & Ermon (2020).

B.3. DGM metrics.

Justification of metrics providing a reasonable and consistent evaluation of images synthesised by generative models is far from a solved problem. Current popular metrics often make use of the final or penultimate layer activations of a heavily-trained convolutional network, such as the Inception v3 model of Szegedy et al. (2015). Despite obvious bias toward generative models trained on similar datasets and in similar manners, as well as the plethora of more performant models since Inception v3 was trained in 2015, these metrics persist, should be used for comparison with the literature, and are now described.

When comparing generated samples to one another, the *Inception score* (Salimans et al., 2016)

$$\text{IS} = \exp(\mathbb{E}_{\mathbf{x}\sim\mathcal{G}}[\text{KL}(p(\mathbf{y}|\mathbf{x}) \parallel p(\mathbf{y}))]), \tag{96}$$

intuitively rewards low entropy classification of generated samples ($\mathbf{x} \sim \mathcal{G}$), as well as variation. Alternatively, the popular *Fréchet inception distance* (FID, (Heusel et al., 2017)) compares Inception v3 activation statistics between generated samples and the samples used to train the generative model, requiring thousands of new samples at evaluation time.

$$\text{FID} = \|\mu_{\mathcal{R}} - \mu_{\mathcal{G}}\|_2^2 + \text{tr}\left(\Sigma_{\mathcal{R}} + \Sigma_{\mathcal{G}} - 2(\Sigma_{\mathcal{R}}\Sigma_{\mathcal{G}})^{1/2}\right), \tag{97}$$

where these values are activation statistics and $\mathcal{R}$ refers to the real dataset. The FID approach was taken further by *kernel inception distance* (KID, (Bińkowski et al., 2018))

$$k(\mathbf{x}_1, \mathbf{x}_2) = \left(\mathbf{x}_1^T\mathbf{x}_2/n + 1\right)^3 \tag{98}$$

$$K(\mathbf{x}_1, \mathbf{x}_2) = k(\phi_{\text{I-v3}}(\mathbf{x}_1), \phi_{\text{I-v3}}(\mathbf{x}_2)), \tag{99}$$

where $\phi_{\text{I-v3}}(\cdot)$ maps to an nD Inception v3 layer, the cubic exponent accounts for skew, crucially no parametric form for the distribution is assumed, and an average over all real-fake pairs is taken.

An alternative approach to sample quality assessment is to directly calculate distribution overlap. In Sajjadi et al. (2018), the authors used local n-balls to form high-dimensional equivalents of *precision* and *recall*, avoiding pathological examples of models with equal FID but visually juxtaposed sample quality. This idea was later extended to the more localised and precise *density* and *coverage* metrics of Naeem et al. (2020), where neighbourhoods are instead built from the k nearest neighbours. The mathematical definitions of these concepts can be found in Naeem et al. (2020).

B.4. Additional figures

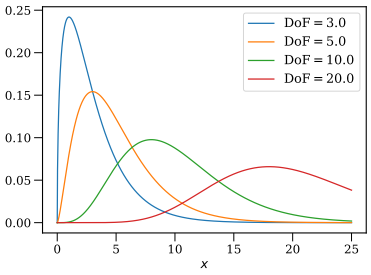
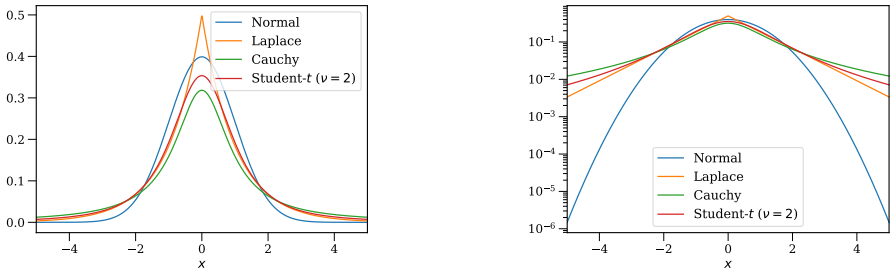


Figure 15. The chi-squared distribution for different degrees of freedom (DoF).



(a) Linear scale. (b) Log scale.

Figure 16. Comparison of the Gaussian distribution and common heavy tailed distributions.

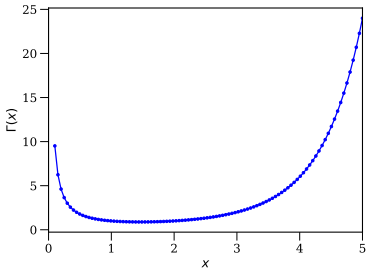
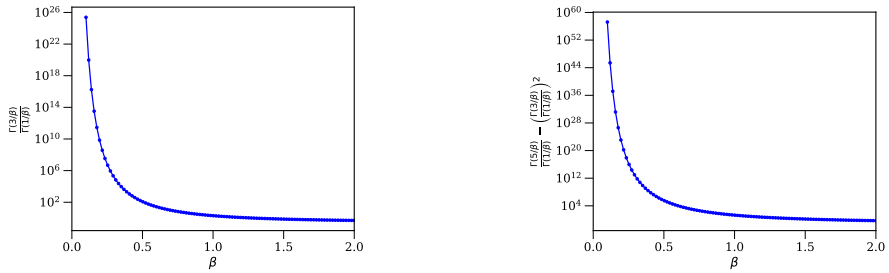


Figure 17. $\Gamma(x)$ in the relevant range for generalised normal noise, $x \in [0.1, 5]$.



(a) Mean scaling factor from (30). (b) Variance scaling factor from (31).

Figure 18. Scaling factor trends against β.