Score-Based Models Detect Manifolds

Anonymous Author(s)
Affiliation
Address
email

Abstract

Score-based generative models (SGMs) need to approximate the scores $\nabla \log p_t$ of the intermediate distributions as well as the final distribution p_T of the forward process. The theoretical underpinnings of the effects of these approximations are still lacking. We find precise conditions under which SGMs are able to produce samples from an underlying (low-dimensional) *data manifold* $\mathcal{M}$. This assures us that SGMs are able to generate the "right kind of samples". For example, taking $\mathcal{M}$ to be the subset of images of faces, we find conditions under which the SGM robustly produces an image of a face, even though the relative frequencies of these images might not accurately represent the true data generating distribution. Moreover, this analysis is a first step towards understanding the generalization properties of SGMs: Taking $\mathcal{M}$ to be the set of all training samples, our results provide a precise description of when the SGM memorizes its training data.

1 Introduction

Score-based generative models ([Sohl-Dickstein et al., 2015, Song and Ermon, 2019, Song et al., 2021, Vahdat et al., 2021]) and the related models ([Bordes et al., 2017, Ho et al., 2020, Kingma et al., 2021]) have shown great empirical success in many areas, such as image generation ([Jolicoeur-Martineau et al., 2021, Nichol and Dhariwal, 2021, Dhariwal and Nichol, 2021, Ho et al., 2022]), audio generation ([Chen et al., 2021, Kong et al., 2021, Jeong et al., 2021, Popov et al., 2021]) as well as in other applications ([De Bortoli et al., 2021, Zhou et al., 2021, Cai et al., 2020, Luo and Hu, 2021, Meng et al., 2021, Saharia et al., 2021, Li et al., 2022, Sasaki et al., 2021]). Recently some progress has been made to bridge the gap between the different approaches ([Song et al., 2021, Huang et al., 2021]) through the framework of SDEs and reverse SDEs.

Score-based generative models (SGMs) are set up to sample from a measure of interest on $\mathbb{R}^d$. Henceforth we will refer to this target measure as μ_{data}, denoting its support by $\mathcal{M}$. In practical situations $\mathcal{M}$ is often a low-dimensional substructure of the full space according to the *manifold hypothesis* ([Bengio et al., 2013, Pope et al., 2021]). When implementing SGMs one needs to make numerical approximations, and therefore the SGM will generate samples from μ_{data} only approximately. Instead, the samples will stem from a proxy measure $\mu_{\text{sample}} \approx \mu_{\text{data}}$. The performance of SGMs can then be measured by comparing μ_{data} to μ_{sample}. We will see under which conditions μ_{data} and μ_{sample} are equivalent (see Appendix A.1 for a definition), and in particular are supported on the same set $\mathcal{M}$. Therefore, even if μ_{sample} might not perfectly reproduce μ_{data}, it will still produce the "right kind of samples". We also find necessary conditions for SGMs to be able to generalize and produce novel samples instead of memorizing the training data. Strikingly, numerical error in the approximation of the score function turns out to be indispensable for any generalization beyond the training set to occur.

SGMs first diffuse each data sample x_0 by gradually adding noise, leading to a path (x_t), $t \in [0, T]$. This process corresponds to transforming the initial distribution $p_0 = \mu_{\text{data}}$, from which the training data stems, into a noisy (final) distribution p_T through a sequence $(p_t)_{t\in[0,T]}$. The final distribution

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

p_T has had so much noise injected that not much of the original signal μ_{data} is left. We study the case in which the forward process can be expressed in terms of a Stochastic Differential Equation (SDE), referred to as the *forward SDE*:

$$\begin{array}{rcl} \mathrm{d}X_t & = & \beta(X_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \\ X_0 & \sim & \mu_{\text{data}}. \end{array} \tag{1}$$

We denote by p_t the density of X_t. Corresponding to (1) we define the *reverse SDE*:

$$\begin{array}{rcl} \mathrm{d}Y_t & = & -\beta(Y_t)\mathrm{d}t + \sigma\sigma^T \nabla \log p_{T-t}(Y_t)\mathrm{d}t + \sigma \mathrm{d}B_t. \\ Y_0 & \sim & q_0, \end{array} \tag{2}$$

We refer to the density of Y_t as q_t. This reverse SDE has the property that if q_0 is chosen to be p_T (the final distribution induced by (1)), then it transforms the distributions backwards, $q_t = p_{T-t}$, in particular implying $q_T = p_0 = \mu_{\text{data}}$. This provides us with a mean to sample from μ_{data}: We first sample from p_T and then run (2) on those samples, transforming the samples to samples from μ_{data}. Note that if $\mathcal{M}$ is a low-dimensional substructure of the space, the density p_0 does not exist, and $\nabla \log p_{T-t}$ explodes as $t \to T$. Therefore the above SDE only makes sense on $[0, T)$ and it is not clear in which sense Y_T exists a priori. For the reverse SDE to be well defined on $[0, T)$ we have to make some assumptions, see Assumption 1. We will prove in Section 2.2 that these Assumptions hold for all popular SDEs used in SGMs.

There are two necessary approximations when implementing a SGM. First, we have no access to p_T and therefore need to replace it by a proxy distribution $q_0 = \mu_{\text{prior}} \approx p_T$. Second, the scores $\nabla \log p_t$ are not available and need to be replaced by a neural net trained via score-matching techniques ([Vincent, 2011, Hyvärinen and Dayan, 2005]). Therefore the sample distribution $\mu_{\text{sample}} = q_T$ will in general differ from μ_{data}. We study the distance of q_T to μ_{data}. In particular, we show under which circumstances q_T will be supported on $\mathcal{M}$.

We first study how the error made while approximating p_T propagates to q_T.

Theorem 1 *Assume that Assumption 1 (see below) holds and that μ_{prior} is absolutely continuous with respect to p_T. Then the following hold.*

(i) Let Y_t be a solution to (2) *on* $[0, T)$. *The limit $Y_T := \lim_{t\to T} Y_T$ exists almost surely. We refer to its distribution by μ_{sample}. The distribution μ_{sample} is absolutely continuous with respect to μ_{data}. If p_T and μ_{prior} are equivalent, then so are μ_{sample} and μ_{data}.*

(ii) Furthermore, for any f-Divergence D_f,

$$D_f(\mu_{sample}|\mu_{data}) \leq D_f(\mu_{prior}|p_T) \quad \textit{and} \quad D_f(\mu_{data}|\mu_{sample}) \leq D_f(p_T|\mu_{prior}).$$

Some choices for D_f are for example the KL-divergence or the total variation distance. In Section 3 we will explicitly bound the KL-divergence for some common SGM implementations. Next we investigate how the approximation of $\nabla \log p_t$ impacts μ_{sample}.

Theorem 2 *Denote by $s(x, t)$ an approximation to $\nabla \log p_{T-t}$ and define the approximation error $e(x, t) = \nabla \log p_{T-t} - s(x, t)$. Assume that Assumption 2 holds (see below).*

Let $\tilde{Y}_t$ be a solution to

$$\begin{array}{rcl} \mathrm{d}\tilde{Y}_t & = & -\beta(\tilde{Y}_t)\mathrm{d}t + \sigma\sigma^T s(\tilde{Y}_t, t)\mathrm{d}t + \sigma \mathrm{d}B_t, \\ \tilde{Y}_0 & \sim & q_0 \end{array} \tag{3}$$

on $[0, T)$. *Then $\tilde{Y}_T = \lim_{t\to T} \tilde{Y}_t$ is well defined. Moreover, its distribution is equivalent to the distribution of Y_T.*

Putting the theorems together, we see that even if we start in the wrong initial distribution $\mu_{\text{prior}} \neq p_T$ and then make an error while approximating the drift $\nabla \log p_t$, the sample measure μ_{sample} will still be equivalent to μ_{data}. Loosely speaking, if $\mathcal{M}$ is the set of all high resolution pictures of faces, then μ_{sample} will also generate samples from that set, just maybe not in the same relative frequency as μ_{data}.

In practice however we can only access μ_{data} through samples $x_i \sim \mu_{\text{data}}$, $i = 1, \ldots, N$. Taking $\mathcal{M} = \{x_i\}_{i=1}^N$ (see Section 4.2), we see that for an SGM to produce novel samples, one of the

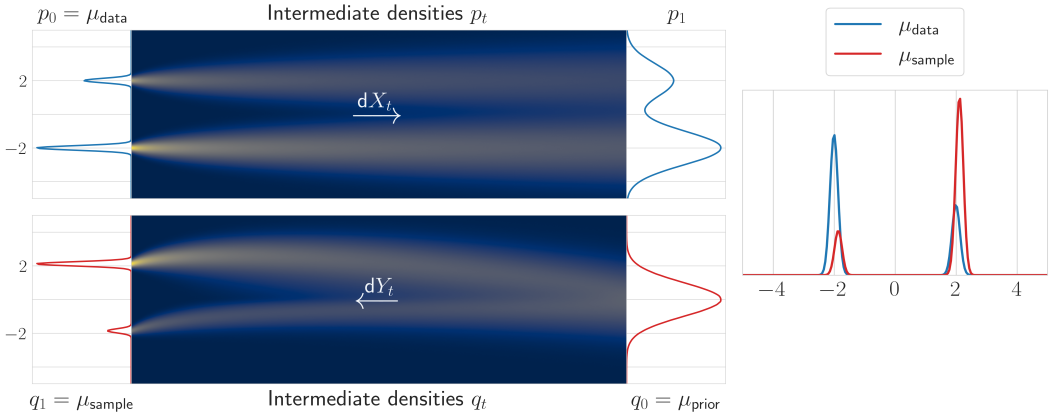

Figure 1: *Top left*: The leftmost plot shows the true data distribution μ_{data}, i.e. a Gaussian mixture. The heat maps show the intermediate densities p_t of X_t, followed by line plots of p_t for $t = 1$.
Bottom left: The rightmost plot shows μ_{prior}, which is a standard Gaussian and differs from p_1. We start the reverse SDE (2) in μ_{prior}. But instead of using the real score, we use an approximation $e(x,t) = \nabla \log p_{T-1} + 1$ with a constant error of 1. Again, the heat maps show how the densities q_t of Y_t evolve backwards in time. The leftmost plot shows the resulting distribution q_t which is equal to the sample distribution μ_{sample}.
Right: The densities μ_{data} and μ_{sample} are shown for direct comparison. We see that the approximation errors in μ_{prior} and the drift lead to a wrong sample distribution $\mu_{\text{sample}} \neq \mu_{\text{data}}$. Nevertheless, μ_{sample} is supported in the same area as μ_{data}. For details on the numerical implementation see Appendix B.

assumptions in Theorem 1 and 2 has to be violated. The Assumptions of Theorem 1 will usually be fulfilled, see Section 2.2.

Therefore, we can conclude that Assumption 2 has to be violated for generalization, and indeed this is expected to be the case in realistic scenarios. If the data is supported on a low-dimensional substructure, $\nabla \log p_{T-t}$ will explode almost everywhere as $t \to T$. For that reason the approximation error e will to grow to infinity as $t \to T$ (see also Section 4.1). Interestingly, this approximation error is necessary for the SGMs to generalize and produce novel samples $x \notin \{x_i\}_{i=1}^N$. The theorems are illustrated in Figure 1.

We now proceed as follows. In Section 2.1 we will state the most commonly used forward SDEs and their properties and show that they fulfil Assumption 1. In Section 5 we prove Theorem 1 and Theorem 2. In Section 3 we discuss how well p_T is approximated by μ_{prior} for commonly used SGD implementations. In Section 4.2 we then discuss how $s(x,t)$ is trained and in which way it approximates $\nabla \log p_t$.

2 Popular SDEs satisfy the assumptions

In this section we will study a few SDEs used in practice and show that they satisfy the following Assumption.

Assumption 1 *Let C and $C_{S,N}$ be constants that can change from appearance to appearance.*

(i) β is globally Lipschitz, i.e. $\|\beta(x) - \beta(y)\| \leq C\|x - y\|$.

(ii) β grows at most linearly, i.e. $\|\beta(x)\| \leq C(1 + \|x\|)$

(iii) X_t has a density $p_t \in C^1$ for every $t > 0$ and $\int_{t_0}^1 \int_{\|x\|<R} |p_t(x)|^2 + \|\nabla_x p_t(x)\|^2 \mathrm{d}x\mathrm{d}t < \infty$ *for any $R > 0$ and $0 < t_0 \leq T$.*

Furthermore, for each $S \in (0,T)$,

(iv) $\nabla \log p_t$ *is locally Lipschitz,* $\|\nabla \log p_t(x) - \nabla \log p_t(y)\| \leq C_{S,N}\|x-y\|$ *for all* $t \in [S,T]$ *and* $\|x\|, \|y\| \leq N$.

Conditions (i)-(iii) are technical conditions on the forward SDE. They ensure that if we run X_t backwards in time, $X_t^R := X_{T-t}$ will be a solution to the reverse SDE (2) on $[0,T)$. The last condition then ensures that the solutions to the reverse SDE are unique, therefore we will be able to transmit the properties of X^R to any other solution Y_t of (2).

2.1 Popular Methods and their Respective SDEs

The first works on SGMs studied discrete forward and backward processes. Nevertheless, the transition kernels and algorithms proposed in those works can be seen as discretisations of some well-known SDEs. More recent works have studied this connection and state the algorithms in terms of SDEs ([Song et al., 2021, Huang et al., 2021]).

Brownian Motion: The works [Song and Ermon, 2019, 2020] can be seen as a discretization of the SDE

$$\mathrm{d}X_t = \sigma(t)\mathrm{d}W_t.$$

Denoting $h(t) = \int_0^t \sigma(s)\mathrm{d}s$, the solution to the above process can be explicitly stated as a time-changed Brownian motion, $X_t = W_{h(t)}$. The time-change can help in the implementation but does not alter the qualitative behaviour of the reverse SDE. In our following analysis we therefore set $\sigma(t) = 1$. Nevertheless, our results still hold for any positive $\sigma(t)$.

Ornstein-Uhlenbeck Process: The works [Sohl-Dickstein et al., 2015, Ho et al., 2020] can be seen as a discretization of

$$\mathrm{d}X_t = -\frac{1}{2}\alpha(t)X_t\mathrm{d}t + \sqrt{\alpha(t)}\mathrm{d}W_t,$$

which is an Ornstein-Uhlenbeck process. Again, the parameters α_t are a time-change and do not influence the properties that we are investigating in this paper. Therefore, to simplify notation, we again set $\alpha(t) = 1$.

Critically Damped Langevin Dynamics (CLD): The work [Dockhorn et al., 2021] studies a second order SDE. Here artificial velocity coordinates V_t are introduced and the system under consideration is

$$\begin{aligned} \mathrm{d}X_t &= V_t, \\ \mathrm{d}V_t &= -X_t - 2V_t + 2\mathrm{d}W_t, \end{aligned}$$

where $X_0 \sim p_{\text{data}}$, $V_0 \sim \mathcal{N}(0, I)$. For generation one runs the reverse SDE in X_t and V_t but discards the V coordinate at the end. The work [Dockhorn et al., 2021] also includes the parameters M and γ. We set both to 1 as they do in their numerical experiments.

2.2 Proving that the assumptions hold

We will now prove that for all SDEs considered in Section 2.1, Assumption 1 holds. To simplify the calculations we assume that the data manifold $\mathcal{M} = \text{supp}(\mu_{\text{data}})$ is contained in a ball of diameter M. This is a natural assumption for many data sets. Nevertheless, we note that this assumption could be weakened by additional technical effort.

First, we note that all SDEs in Section 2.1 are linear SDEs. Therefore, their transition kernels are Gaussian ([Pavliotis, 2014, Section 3.7]):

$$\pi_t(Z_t = z | Z_0 = z_0) = \mathcal{N}(z; m_t(z_0), \Sigma_t).$$

The explicit form of m_t and Σ_t differ for each of the SDEs and can be found in Appendix C.1. We remark that Σ_t does not depend on the initial condition z_0. The transition kernel above gives us the distribution of the SDE started in a single point z_0. Since we start the SDE in μ_{data} we need to average over μ_{data} to get the marginal at time t:

$$p_t(z) = \int_{\mathbb{R}^d} \mathcal{N}(z; m_t(z_0), \Sigma_t)\mu_{\text{data}}(z_0)\mathrm{d}z_0. \tag{4}$$

We can also compute the additional drift in the reverse SDE (see Appendix C.2),

$$\nabla \log p_t = \frac{\nabla p_t(z)}{p_z(z)} = \Sigma_t^{-1}(z - \mathbb{E}[m_t(Z_0)|Z_t = z]). \tag{5}$$

The drift is intuitive, especially for the Brownian Motion, where $m_t(z) = z$ and $\Sigma_t = t$. In that case, $\mathbb{E}[Z_0|Z_t = z]$ is the expected position from which the SDE was started on the manifold, if it is at position z at time t. For a point z which has distance $\delta > 0$ to $\mathcal{M}$, $\|z - \mathbb{E}[Z_0|Z_t = z]\| > \delta$, since Z_0 is in $\mathcal{M}$ almost surely. Therefore, we can infer that $\nabla \log p_t(z)$ explodes like $\frac{\delta}{t}$ as $t \to 0$. A similar analysis can be conducted for the other two SDEs using the formulas for the transition kernels in Appendix C.1. We are now ready to prove the following

Lemma 1 *Assume that the data manifold $\mathcal{M}$ is contained in a ball of radius M. Then all the methods introduced in Section 2.1 fulfil Assumption 1.*

Proof The forward drifts are $\beta(x) = 0$, $\beta(x) = -\frac{\alpha}{2}x$ and $\beta(x, v) = (v, -x - 2v)$ for the Brownian Motion, the OU-Process and the Critically Damped Langevin Dynamics respectively. In particular, these are all linear maps and therefore fulfil conditions (i) and (ii) of Assumption 1.

We show in Appendix C.1 that $\log p_t$ is C^∞ in t and x for $t > 0$. Therefore we can integrate p_t and its derivative over compact sets, implying that condition (iii) holds. Furthermore, the Hessian w.r.t. (x, t) is continuous and obtains its maximum and minimum on the compact set $[S, T] \times B_N$, where B_N is the ball of diameter N around the origin. Therefore the gradient $\nabla \log p_t$ is Lipschitz on $[S, T] \times B_N$, which proves (iv). □

3 Distance from p_T to μ_{prior}

We have seen in Theorem 1 that the distance between μ_{sample} and μ_{data} is directly related to the distance between p_T and μ_{prior}. The OU-process and the CLD are well-known SDEs in the Markov Chain Monte Carlo community. They form building blocks of algorithms like Langevin MCMC ([Roberts and Tweedie, 1996, Welling and Teh, 2011]) or Hamiltonian Monte Carlo ([Neal et al., 2011]) and the speed at which they converge to their stationary distribution is well studied ([Roberts and Tweedie, 1996, Markowich and Villani, 2000, Bou-Rabee et al., 2010, Eberle et al., 2019, Cao et al., 2019]). It is known that their intermediate distributions p_t will converge to $\mu_{\text{prior}} = \mathcal{N}(0, I_d)$ at an exponential rate.

The Brownian motion however does not converge to a stationary distribution and therefore one has to choose a different μ^T_{prior} for each T to approximate p_T. We now study the distance of μ^T_{prior} to p_T. If we transfer the suggestions of [Song et al., 2021, Appendix C] to our setting, they propose to set μ^T_{prior} to $\mathcal{N}(0, T)$. However, p_t has the same mean as μ_{data} and has variance $\text{Var}[\mu_{\text{data}}] + T$. Loosely speaking, in contrast to the other two SDEs, the Brownian Motion does not forget its initial mean and covariance. As we see in the following lemma, we can decrease the distance between p_T and μ^T_{prior} by choosing the mean and covariance of μ^T_{prior} equal to the mean and covariance of p_T.

Lemma 2 *Let p_T be the T-time marginal of the Brownian motion process X_t of Section 2.1. The following minimization problem*

$$\min_{m,C} KL(p_T \mid \mathcal{N}(m, C))$$

is minimized by $m_T = \mathbb{E}[\mu_{data}]$ and $C_T = \text{Cov}[\mu_{data}] + TI_d$. If we restrict the covariance to be a multiple of the identity matrix, i.e.

$$\min_{m,c} KL(p_T \mid \mathcal{N}(m, cI_d)),$$

the problem is solved by choosing m as above and c as $c = \mathbb{E}[\|X_t - m\|^2] = trace(C_T)$.

This result is a slight variation on the well known fact that the KL-projection in the second argument matches its moment. We prove it in Appendix E. Equivalently to choosing the mean of μ^T_{prior} equal to the mean of p_T, $\mathbb{E}[p_T] = \mathbb{E}[p_0]$, one can normalize the training data to have mean 0 and then also centre μ^T_{prior}. However, the covariance matrix will not be equal to I_d after common normalization methods.

The next lemma bounds the distance of p_T and μ^T_{prior} in the Brownian motion case.

Lemma 3 *Let p_t be the time t-marginal of a Brownian motion with initial condition μ_{data}. Denote by $c_i, i = 1, \dots, d$, the eigenvalues of $\mathrm{Cov}(\mu_{data})$. Let μ^T_{prior} be any normal distribution. Then*

$$KL(p_T|\mu^T_{prior}) \leq \frac{1}{2} \log \left(\frac{\prod_{i=1}^d (c_i + T)}{T^d} \right).$$

This lemma is proven by explicitly bounding the distance of p_T to the optimal normal distribution from Lemma 2, see Appendix E.

4 Approximation of the score $\nabla \log p_t$

4.1 Violation of the assumption

First, we state the assumption that the approximation has to fulfil in order for Theorem 2 to be valid.

Assumption 2 *Let $s(x,t), e(x,t)$ be defined as in Theorem 2. We assume that the reverse SDE $\tilde{Y}_t$ has a solution on $[0,T)$. For $t < T$, we define the Girsanov weights*

$$Z_t = \exp \left(\int_0^t \sigma^T e(s, Y_s) \cdot \mathrm{d}B_s - \frac{1}{2} \int_0^t \|\sigma^T e(s, Y_s)\|^2 \mathrm{d}s \right)$$

and assume that the Z_t are a uniformly integrable martingale (see Appendix A.2).

Furthermore, we assume that the approximate drift $s(x,t)$ also fulfils condition (iv) of 1.

We again need (iv) to hold to have uniqueness in law of the reverse SDE, see the discussion after Assumption 1. However, the local Lipschitzness is implied if $s(x,t)$ is C^2, which we can expect in practice.

We already stated in Theorem 2 that a sufficient condition for Assumption 2 to hold is for e to be bounded. From Section 2.2 we know that for each fixed z outside $\mathcal{M}$, $\nabla \log p_t(z)$ will explode. Since s is a numerical approximation to $\nabla \log p_t$, it will be finite. Therefore e will also grow to infinity. Therefore, if $\mathcal{M}$ is a lower-dimensional substructure, we cannot expect e to be bounded.

Note however, that a global bound on e is stronger than what we need for Assumption 2. In particular, the approximation error e is only evaluated along the path Y_s. If we follow the intuition of Section 2.2 in the Brownian motion case, $\nabla \log p_{T-t}(Y_t)$ would approximately have magnitude $\frac{\mathrm{dist}(Y_t, \mathcal{M})}{T-t}$. If Y_t goes to $\mathcal{M}$ fast enough, then $\nabla \log p_t$ may stay bounded along the typical paths of Y.

Let us briefly investigate this. Continuing with the Brownian motion case, if we neglect the approximation errors, we can think of Y_t as $X_{T-t} = X_0 + \sqrt{T-t}Z$. Here X_0 is distributed according to μ_{data} and lies on $\mathcal{M}$ and Z is standard normal. Especially, $\mathbb{E}[\|\sqrt{T-t}Z\|] = \sqrt{T-t}\,\mathbb{E}[\|\sqrt{T-t}Z\|]$ will be of magnitude $\sqrt{T-t}$. Therefore, we can expect the distance of Y_t to $\mathcal{M}$ to be of order $\sqrt{T-t}$. Because of that we can expect $\nabla \log p_{T-t}(Y_t)$ to have magnitude $\frac{1}{\sqrt{t}}$ along a typical path of Y_t. This is unbounded and the approximation error will tend to infinity as $t \to T$.

4.2 Score approximation with a finite number of samples

The score $\nabla \log p_t$ is approximated by minimizing a variant of

$$L(\theta, t) = \mathbb{E}_{p_t(x)}[\|\nabla \log p_t(x) - s_\theta(x,t)\|]. \tag{6}$$

for $t \in (0, T]$. This is done via score matching techniques (see [Hyvärinen and Dayan, 2005, Vincent, 2011, Song et al., 2020]). The above expectation is taken over p_t, which itself depends on μ_{data},

$$p_t(x) = \mathbb{E}_{\mu_{\text{data}}(x_0)}[\pi_{t|0}(x|x_0)].$$

In practice we do not have access to the full data distribution μ_{data}. Therefore we replace μ_{data} by a its empirical measure $\hat{\mu}_{\text{data}}$, which is supported on the training set,

$$\hat{\mu}_{\text{data}}(x) = \text{Unif}\{x_1, \dots, x_n\}. \tag{7}$$

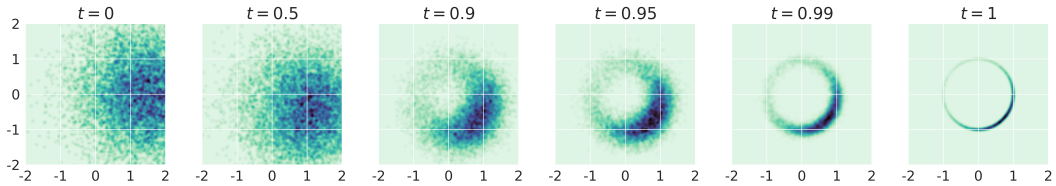

Figure 2: Here μ_{data} is chosen as the uniform distribution on the unit sphere $\mathcal{S}^1$. We use the Brownian motion for the forward SDE. We can compute the exact score $\nabla \log p_t(x)$. We disturb it with the vector $v = (x = 0, y = -1)$ and define the approximation $s(x,t) = \nabla \log p_t(x) + v$. Furthermore, we purposely adopt a poor approximation $\mu_{\text{prior}} \not\approx p_1$ by setting $\mu_{\text{prior}} = \mathcal{N}(m, I_2)$, where $m = (x = 1.5, y = 0)$. We then run the reverse SDE (3). The figure shows heat maps of the intermediate distributions q_t of the reverse SDE. At time $t = 1$ we reach $q_1 = \mu_{\text{sample}}$. We see that μ_{sample} is a distribution on $\mathcal{M} = \mathcal{S}^1$, albeit not the uniform one.

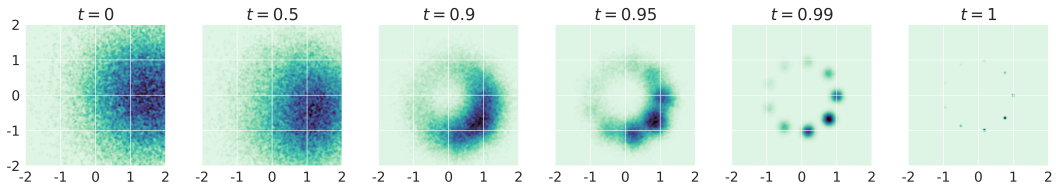

Figure 3: Here the numerical experiment of Figure 2 is repeated but with μ_{data} chosen to be the uniform distribution on $\mathcal{M} = \{x_i\}_{i=1}^9$, where the x_i are 9 evenly spaced points on the unit sphere $\mathcal{S}^1$. Notice that while the approximation errors cause a non-uniform distribution the training examples, μ_{sample} is still supported solely on $\mathcal{M}$ and will not generate novel samples.

This leads to a Monte-Carlo estimate of $\hat{p}_t$ of p_t,

$$\hat{p}_t(x) = \mathbb{E}_{\hat{p}_{\text{data}}(x_0)}[p_{t|0}(x|x_0)] = \frac{1}{N}\sum_{i=1}^{N} p_{t|0}(x|x^i) \approx p_t(x).$$

The surrogate loss

$$\hat{L}(\theta, t) = \mathbb{E}_{\hat{p}_t(x)}[\|\nabla \log \hat{p}_t(x) - s_\theta(x,t)\|], \tag{8}$$

can be evaluated and used for training. Instead of interpreting this loss as an approximation to (6), we can also interpret it as the exact loss for the distribution $\hat{\mu}_{\text{data}}$. However in this case, $\mathcal{M} = \text{supp}(\hat{\mu}_{\text{data}}) = \{x_1, \ldots, x_n\}$. If Theorem 2 would hold and μ_{sample} would be equivalent to μ_{data}, then μ_{sample} could only draw samples from $\mathcal{M}$. In other words, the SGM would have memorized the training set. This is also visualized in Figures 2 and 3.

However, empirically it has been shown that SGMs are able to create novel samples (see, for example [Dhariwal and Nichol, 2021]). Furthermore, in the case of image generation, the images are sharp. In particular, they probably come from the same low-dimensional substructure that the training samples of μ_{data} are from. Therefore the approximation errors made by s_θ are non-Gaussian and probably tangent to $\mathcal{M}$. The form and properties of the errors e are closely related to the ability of the SGM to generalize.

5 Proofs of the theorems

We now give proofs of our main results and briefly summarize the key steps in an intuitive way. In our study we would like to include the case when μ_{data} is degenerate and supported on a low-dimensional substructure $\mathcal{M}$. As we have seen in Section 2.2, this can lead to an exploding drift in the reverse SDE as $t \to T$. Nevertheless, in order to understand the properties of μ_{data}, it is crucial to study the properties of solutions to the reverse SDE at time $t = T$. This is where the main mathematical difficulties come from. The proofs are mostly independent of the specific form of the forward SDE and hold for more general forward/backward SDEs than those stated in Section 1, see Appendix D.

5.1 Theorem 1

We now proceed with proving Theorem 1.

Proof Let P be the measure on $\Omega = C([0,T],\mathbb{R}^n)$ induced by the forward SDE (1). P has marginals p_t. Denote by X_t the canonical projections $X_t(\omega) = \omega(t)$ for $\omega \in \Omega$. We define Q through

$$\frac{\mathrm{d}Q}{\mathrm{d}P}(\omega) = \frac{\mathrm{d}\mu_{\text{prior}}}{\mathrm{d}p_T}(\omega(T)).$$

By the data processing inequality we obtain (see [Liese and Vajda, 2006, Theorem 14]),

$$KL(q_T|\mu_{\text{data}}) \leq KL(Q|P) = KL(\mu_{\text{prior}}|p_T). \tag{9}$$

It remains to prove that by running Q backwards we obtain a solution to (2) started in μ_{prior}. We denote the generator of the reverse SDE (2) by $\mathcal{L}$. Denote by Q^R and P^R the time reversals of Q and P. Our assumption are such that P^R is a Markov process solving the martingale problem for $\mathcal{L}$ (see [Haussmann and Pardoux, 1986, Theorem 2.1]). A short calculation shows that Q^R is still Markov (see, for example [Léonard, 2011, Proposition 4.2]). Furthermore for $f \in C_c^\infty(\mathbb{R}^n)$,

$$\begin{aligned}
&\mathbb{E}_{Q^R}\left[f(X_t) - f(X_s) - \int_s^t \mathcal{L}f(X_r)\mathrm{d}r|X_s\right] \\
&= \frac{\mathbb{E}_{P^R}\left[\left(f(X_t) - f(X_s) - \int_s^t \mathcal{L}f(X_r)\mathrm{d}r\right)\frac{\mathrm{d}\mu_{\text{prior}}}{\mathrm{d}p_T}(X_0)|X_s\right]}{\mathbb{E}_{P^R}\left[\frac{\mathrm{d}\mu_{\text{prior}}}{\mathrm{d}p_T}(X_0)|X_s\right]} \\
&= \frac{\mathbb{E}_{P^R}\left[\left(f(X_t) - f(X_s) - \int_s^t \mathcal{L}f(X_r)\mathrm{d}r\right)|X_s\right]\mathbb{E}_{P^R}\left[\frac{\mathrm{d}\mu_{\text{prior}}}{\mathrm{d}p_T}(X_0)|X_s\right]}{\mathbb{E}_{P^R}\left[\frac{\mathrm{d}\mu_{\text{prior}}}{\mathrm{d}p_T}(X_0)|X_s\right]} \\
&= \mathbb{E}_{P^R}\left[\left(f(X_t) - f(X_s) - \int_s^t \mathcal{L}f(X_r)\mathrm{d}r\right)|X_s\right] = 0.
\end{aligned}$$

In the second equality we used the Markov property of P^R. In the last one we used that P^R solves the martingale problem for $\mathcal{L}$. Therefore also Q^R solves the martingale problem for $\mathcal{L}$. Denote by Y_t a solution to (2) on $[0,T)$. Since solutions to (2) are unique in law on $[0,S]$ for $S < T$ (see [Karatzas and Shreve, 2012, Section 5.2]) and the solutions are continuous, the law of Y is equal to Q^R on $[0,T)$. But the paths of Q^R are continuous on $[0,T]$. Therefore, Y can be extended to $[0,T]$, i.e. the limit $Y_T := \lim_{t\to T} Y_t$ exists almost surely and its distribution is equal to the T-time marginal distribution of Q^R, which is the 0-time marginal of Q. We denote the marginals of Q and P by Q_t and P_t. Since Q is absolutely continuous with respect to P, $Q_0 = \mu_{\text{sample}}$ is absolutely continuous with respect to $P_0 = \mu_{\text{data}}$. Analogously, if μ_{prior} and p_T are equivalent, then so are P and Q and therefore P_0 and Q_0. This proves (i).

(ii) is a consequence of the data processing inequality for f-divergences ([Liese and Vajda, 2006, Theorem 14]), analogous to (9). □

The main idea of this proof is that we look at the forward SDE for X_t first. It induces a distribution $\mathbb{P}$ over all continuous paths in $\Omega = C([0,T],\mathbb{R}^d)$. If we reverse the time direction of this distribution on Ω, we get a solution to the reverse SDE, started in p_T. This reverse solution is well behaved as $t \to T$, since $Y_T = X_0$. The solution for a different initial condition $\mu_{\text{prior}} \neq p_T$ is obtained by reweighting $\mathbb{P}$. This does not change the qualitative behaviour of Y_T, which still exists and is well defined. We then use a uniqueness result to see that any solution of (11) inherits these properties.

5.2 Theorem 2

Proof Denote the space $C([0,T),\mathbb{R}^d)$ by Ω. We also define the natural filtration $\mathcal{F}_t = \sigma(x(s)|s \leq t)$. We denote the distribution of Y on Ω by $\mathbb{P}$. We define $\tilde{\mathbb{P}}$ by reweighting $\mathbb{P}$ with Z_t on $\mathcal{F}_t$. By Girsanov theorem (see [Karatzas and Shreve, 2012, Section 3.5]) we know that the canonical process under $\tilde{\mathbb{P}}$ is a solution to (3) on $[0,T)$. Since Z_t is uniformly integrable, its limit $Z := \lim_{t\to T} Z_t$ exists in L^1.

Furthermore $\tilde{\mathbb{P}}$ is absolutely continuous with respect to $\mathbb{P}$ on $\mathcal{F} = \sigma(\cup_{t<T}\mathcal{F}_t) = \sigma(x(t)|t < T)$ with density Z. We define $x(T) := \lim_{t\to T} x(t)$. Then the event

$$\mathcal{A} = \{x(T) := \lim_{t\to T} x(t) \text{ exists and } x(T) \in \mathcal{M}\}$$

has probability 1 under $\mathbb{P}$ (see Theorem 1) and therefore also under $\tilde{\mathbb{P}}$. Furthermore, $x(T)$ is measurable with respect to $\mathcal{F}$. Therefore the distributions of $x(T)$ under $\mathbb{P}$ and $\tilde{\mathbb{P}}$ are equivalent. The canonical process under $\tilde{\mathbb{P}}$ is therefore a solution of (3), with the property that its time T-marginal is well defined and equivalent to the time T-marginal of (2). We can use uniqueness in law on $[0, S)$ for any $S < T$ and extend it to $[0, T]$ as in the proof of Theorem 1. This shows that every solution to (3) has the desired properties.

Finally we show that if e is bounded, it fulfils Assumptions 2. We define by $H_t = \int_0^t \|e(s, \hat{Y}_s)\| \mathrm{d}s$. Then there is a Brownian motion W_t such that we can write Z_t as

$$Z_t = \exp\left(W_{H_t} - \frac{1}{2}H_t\right).$$

Since e is bounded by M, H_t is bounded by TM. In particular, one can view $Z_t = \mathbb{E}[Z_{TM}|\mathcal{F}_{H_t}]$. Therefore Z_t is uniformly integrable since it can be viewed as a family of conditional expectations. □

Here we essentially applied the Girsanov Theorem on $[0, T)$. Using the uniform integrability of the Girsanov weights Z_t, we are able to extend it to $[0, T]$. Therefore, we can infer that the distribution of Y and $\hat{Y}$ are actually equivalent on the whole path space $C([0, T], \mathbb{R}^d)$. In particular, their time T-marginals will be equivalent too, which is the claim of the theorem.

6 Broader impact

The results deepen the understanding of score-based generative models. As such, they can be seen as a step towards improving the quality of generative models. Therefore the possible negative societal impacts are the same ones that apply to generative modelling in general. First, generative models can be used to create synthetic data that is hard to distinguish from real data (for example images or videos), see [Mirsky and Lee, 2021]. Second, generative models can learn and reproduce biases that are prevalent in the training data ([Esser et al., 2020]). Last, depending on the application, generative models might be used to do creative work that was previously done by humans.

7 Previous studies

The work [De Bortoli et al., 2022] treats a variant of SGMs that is defined on a manifold. In this variant, the full algorithm is defined on a Riemannian manifold $\mathcal{M}$. We study the common SGM implementation under the assumption that the initial distribution is supported on a substructure of $\mathbb{R}^d$. As a matter of fact, we do not require $\mathcal{M}$ to carry any manifold structure in a mathematical sense, since we define it as $\mathcal{M} = \text{supp}(\mu_{\text{data}})$. Theorem 1 in [De Bortoli et al., 2021] studies the distance of the sample measure to the data generating distribution. An explicit rate for the convergence in the total variation distance is shown. However, it is assumed that that the initial distribution has an everywhere positive density, which excludes degenerate μ_{data}. Furthermore, it seems to be challenging to infer equivalence of the measures at any finite time.

8 Conclusion

We conducted a theoretical study of some properties of SGMs. We found explicit conditions under which the sample measure μ_{sample} is equivalent to the true data generating distribution μ_{data}. Under these conditions we can guarantee, that the SGM generates samples that could also be samples from μ_{data}. Furthermore, each sample that can be generated by μ_{data} also has positive probability under μ_{sample}, meaning that the full support is covered.

Since one can not actually access the full support of μ_{data}, but only a finite number of training examples $\{x_i\}_{i=1}^N$, our results can be applied to find conditions under which the SGM memorizes its training data. We believe that this observation provides a first step towards understanding the generalization capabilities of SGMs.

References

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

F. Bordes, S. Honari, and P. Vincent. Learning to generate samples from noise through infusion training. *arXiv preprint arXiv:1703.06975*, 2017.

N. Bou-Rabee, M. Hairer, and E. Vanden-Eijnden. Non-asymptotic mixing of the mala algorithm. *IMA Journal of Numerical Analysis*, 33, 08 2010. doi: 10.1093/imanum/drs003.

R. Cai, G. Yang, H. Averbuch-Elor, Z. Hao, S. J. Belongie, N. Snavely, and B. Hariharan. Learning gradient fields for shape generation. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 364–381. Springer, 2020. doi: 10.1007/978-3-030-58580-8_22. URL `https://doi.org/10.1007/978-3-030-58580-8_22`.

Y. Cao, J. Lu, and L. Wang. On explicit l^2-convergence rate estimate for underdamped langevin dynamics. *arXiv preprint arXiv:1908.04746*, 2019.

N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan. Wavegrad: Estimating gradients for waveform generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=NsMLjcFaO8O`.

M. Costa and T. Cover. On the similarity of the entropy power inequality and the brunn-minkowski inequality (corresp.). *IEEE Transactions on Information Theory*, 30(6):837–839, 1984.

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34, 2021.

V. De Bortoli, E. Mathieu, M. Hutchinson, J. Thornton, Y. W. Teh, and A. Doucet. Riemannian score-based generative modeling. *arXiv preprint arXiv:2202.02763*, 2022.

P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

T. Dockhorn, A. Vahdat, and K. Kreis. Score-based generative modeling with critically-damped langevin diffusion. *CoRR*, abs/2112.07068, 2021. URL `https://arxiv.org/abs/2112.07068`.

A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019.

P. Esser, R. Rombach, and B. Ommer. A note on data biases in generative models. In *NeurIPS 2020 Workshop on Machine Learning for Creativity and Design*, 2020. URL `https://arxiv.org/abs/2012.02516`.

U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205, 1986.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

C.-W. Huang, J. H. Lim, and A. C. Courville. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34, 2021.

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

M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim. Diff-tts: A denoising diffusion model for text-to-speech. In H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, editors, *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3605–3609. ISCA, 2021. doi: 10.21437/Interspeech.2021-469. URL `https://doi.org/10.21437/Interspeech.2021-469`.

A. Jolicoeur-Martineau, R. Piché-Taillefer, I. Mitliagkas, and R. T. des Combes. Adversarial score matching and improved sampling for image generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=eLfqMl3z3lq`.

I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 2012.

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

A. Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=a-xFK8Ymz5J`.

C. Léonard. Stochastic derivatives and generalized h-transforms of markov processes. *arXiv preprint arXiv:1102.3172*, 2011.

H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. doi: 10.1016/j.neucom.2022.01.029. URL `https://doi.org/10.1016/j.neucom.2022.01.029`.

F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

S. Luo and W. Hu. Diffusion probabilistic models for 3d point cloud generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2837–2845. Computer Vision Foundation / IEEE, 2021. URL `https://openaccess.thecvf.com/content/CVPR2021/html/Luo_Diffusion_Probabilistic_Models_for_3D_Point_Cloud_Generation_CVPR_2021_paper.html`.

P. A. Markowich and C. Villani. On the trend to equilibrium for the fokker-planck equation: an interplay between physics and functional analysis. *Mat. Contemp*, 19:1–29, 2000.

C. Meng, Y. Song, J. Song, J. Wu, J. Zhu, and S. Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *CoRR*, abs/2108.01073, 2021. URL `https://arxiv.org/abs/2108.01073`.

Y. Mirsky and W. Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.

R. M. Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11): 2, 2011.

A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

G. A. Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=XJk19XzGq2J`.

V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8599–8608. PMLR, 2021. URL http://proceedings.mlr.press/v139/popov21a.html.

O. Rioul. Information theoretic proofs of entropy power inequalities. *IEEE Transactions on Information Theory*, 57(1):33–55, 2010.

G. O. Roberts and R. L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *CoRR*, abs/2104.07636, 2021. URL https://arxiv.org/abs/2104.07636.

H. Sasaki, C. G. Willcocks, and T. P. Breckon. UNIT-DDPM: unpaired image translation with denoising diffusion probabilistic models. *CoRR*, abs/2104.05358, 2021. URL https://arxiv.org/abs/2104.05358.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11895–11907, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3001ef257407d5a371a96dcd947c7d93-Abstract.html.

Y. Song and S. Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

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

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

A. Vahdat, K. Kreis, and J. Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34, 2021.

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

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

L. Zhou, Y. Du, and J. Wu. 3d shape generation and completion through point-voxel diffusion. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5806–5815. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00577. URL https://doi.org/10.1109/ICCV48922.2021.00577.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The assumptions are followed by a discussion of their strength.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] There are no error to report, we just have small illustrative numerical experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We did not use any existing assets.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use crowdsourcing or conduct research with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]