# Score-Based Generative 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, also called diffusion models ([Sohl-Dickstein et al., 2015, Song and Ermon, 2019, Song et al., 2021b, 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 ([Batzolis et al., 2021, 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., 2021b, Huang et al., 2021]) through the framework of SDEs and reverse SDEs.

In generative modelling one is given samples $\{x^i\}_{i=1}^N$ from a measure $\mu_{\text{data}}$. The task is to learn a measure $\mu_{\text{sample}}$ which approximates $\mu_{\text{data}}$. The performance of a generative model can then be measured by the distance from $\mu_{\text{sample}}$ to $\mu_{\text{data}}$. In practice however, the true data generating distribution $\mu_{\text{data}}$ is unknown. All that is known are the samples $\{x_i\}_{i=1}^n$, which can be used to define the empirical measure $\hat{\mu}_{\text{data}}$,

$$\hat{\mu}_{\text{data}} := \text{Unif}\{x^1, x^2, \dots, x^n\}.$$

Any sample from the empirical measure $\hat{\mu}_{\text{data}}$ will be equal to a training example. Hence, while $\mu_{\text{sample}}$ being close to $\mu_{\text{data}}$ is the final goal, $\mu_{\text{sample}}$ being close to $\hat{\mu}_{\text{data}}$ implies that the generative model has memorized the training data. To summarize, a good generative model will output a measure $\mu_{\text{sample}}$ which is as close to $\mu_{\text{data}}$ as possible, while keeping some distance from $\hat{\mu}_{\text{data}}$, even though it only knows $\mu_{\text{data}}$ through $\hat{\mu}_{\text{data}}$.

Given a target measure $\pi_0$, a score-based generative model (SGM) employs two stochastic differential equations (SDEs). The first one is called the *forward SDE*

$$\begin{aligned} \mathrm{d}X_t &= \beta(X_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \\ X_0 &\sim \pi_0. \end{aligned} \tag{1}$$

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

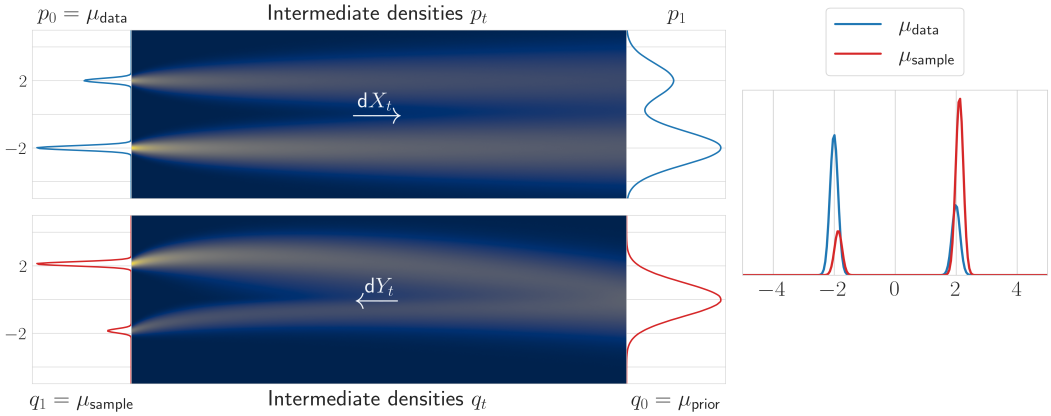



Figure 1: *Top left*: The leftmost plot shows the true data distribution $\mu_{\text{data}}$ which is a Gaussian mixture. The heat maps show the intermediate densities $p_t$ of $X_t$, followed by line plots of $p_t$ for $t = 1$.
*Bottom left*: The rightmost plot shows $\mu_{\text{prior}}$, which is a standard Gaussian and differs from $p_1$. We start the reverse SDE (2) in $\mu_{\text{prior}}$. But instead of using the real score, we introduce an approximation error and use $s(x,t) = \nabla \log p_{1-t}(x) + 3$ with a constant error of 3. Again, the heat maps show how the densities $q_t$ of $Y_t$ evolve backwards in time. The leftmost plot shows the resulting distribution $q_1$ which is used as sample distribution, $\mu_{\text{sample}} = q_1$.
*Right:* The densities $\mu_{\text{data}}$ and $\mu_{\text{sample}}$ are shown for direct comparison. We see that the approximation errors in $\mu_{\text{prior}}$ and the drift lead to an incorrect sample distribution $\mu_{\text{sample}} \neq \mu_{\text{data}}$. Nevertheless, $\mu_{\text{sample}}$ is supported in the same area as $\mu_{\text{data}}$. For details on the numerical implementation see Appendix B.

The marginals of $X_t$ are denoted by $\pi_t$. The forward SDE is run until some terminal time $T$. Furthermore, the *reverse SDE* is defined by

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

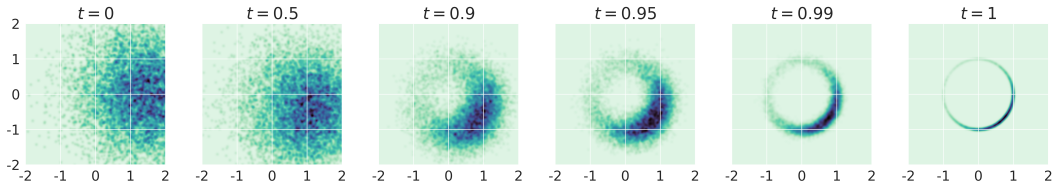
We refer to the marginals of $Y_t$ as $q_t$. The samples are generated from the final distribution $q_T$, i.e. $\mu_{\text{sample}} := q_T$. The reverse SDE has the property that if $q_0$ is chosen to be equal to $\pi_T$, then $q_t = \pi_{T-t}$. In particular, this implies that $q_T = \pi_0$. Therefore, if we have samples from $\pi_T$, we can run the reverse SDE on them to create new samples from $\pi_0$.

In the following we will denote by $p_t$ the marginals of the forward SDE when started in the true data generating distribution $\pi_0 = \mu_{\text{data}}$. We will denote by $\hat{p}_t$ the marginals of the forward SDE when started in $\pi_0 = \hat{\mu}_{\text{data}}$. Optimally, we would like to run the algorithm using $\pi_0 = \mu_{\text{data}}$, i.e. with marginals $\pi_t = p_t$. This is however not possible, since $\mu_{\text{data}}$ itself is unknown.

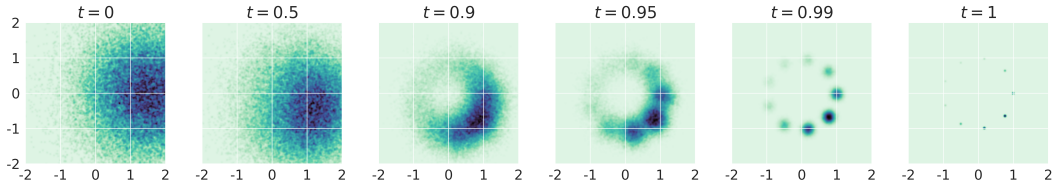
To circumvent the problem of not knowing $p_T$, the forward SDE is chosen such that it forgets its initial condition $p_0$. At time $T$, the marginal $p_T$ is then well approximated by a proxy distribution $\mu_{\text{prior}} \approx p_T$, independently of $p_0$. Additionally, the marginals $p_t$ and therefore the scores $\nabla \log p_t$ cannot be evaluated for $p_0 = \mu_{\text{data}}$. Therefore, the scores are replaced by a neural network $s_\theta(x,t)$, which is trained via score-matching techniques [Vincent, 2011, Hyvärinen and Dayan, 2005].

As a result, SGMs make two approximations. The first one is in approximating $p_T$ by $\mu_{\text{prior}}$. The second one is the approximation of $\nabla \log p_t$ by the neural net $s_\theta(x,t)$. We illustrate this in Figure 1. It is important to understand how these approximations translate to the distance of $\mu_{\text{sample}}$ to $\mu_{\text{data}}$ or $\hat{\mu}_{\text{data}}$.

Some early works already deal with these questions. In [De Bortoli et al., 2021] the total variation distance between $\mu_{\text{sample}}$ and $\mu_{\text{data}}$ in bounded, whereas [Song et al., 2021a] derives bounds with respect to the KL-Divergence. The work [De Bortoli et al., 2021] furthermore derives a result similarly to the second part of Theorem 1, but also treating the errors that are introduced by discretizing the SDE. However, both of these works assume that the initial distribution $\mu_{\text{data}}$ is rather well behaved. In particular, it is assumed that $\mu_{\text{data}}(x) > 0$ for all $x$. We shortly discuss this assumption now.




(a) Here $\mu_{\text{data}}$ 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 perturb 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 $\mu_{\text{prior}} = \mathcal{N}(m, I)$, where $m = (x = 1.5, y = 0)$. We then run the reverse SDE (6). 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 $\mu_{\text{sample}}$ is a distribution on $\mathcal{M} = \mathcal{S}^1$, albeit not the uniform one. Furthermore, we can observe how the errors in the initial conditions and the drift influence the skewed distribution $\mu_{\text{sample}}$. The initial conditions where chosen to have a to large $x$-coordinate on average, whereas the drift was chosen as to prefer lower $y$-coordinates. The distribution $\mu_{\text{sample}}$ is concentrated in areas with high $x$ and low $y$-coordinates.




(b) The above experiment is repeated but with $\mu_{\text{data}}$ 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, $\mu_{\text{sample}}$ is still supported solely on $\mathcal{M}$ and will not generate novel samples.

Figure 2: Perturbing $\nabla \log \hat{p}_t$.

Assuming that $\mu_{\text{data}}(x) > 0$ for any $x$ means that one postulates that every $x$ is a possible sample from $\mu_{\text{data}}$. For example, this implies that even if all samples $\{x_1, \dots, x_n\}$ of $\mu_{\text{data}}$ consist of images of human faces, we say that $\mu_{\text{data}}$ can possibly also generate images of for example furniture, animals or pure white noise. Even though it might have low probability, any combinations of pixels is a possible sample from $\mu_{\text{data}}$. The assumption that $\mu_{\text{data}}$ is actually supported on some lower dimensional substructure $\mathcal{M}$ is well known under the name *manifold hypothesis*, see for example [Bengio et al., 2013, Pope et al., 2021]. In practice this means that $\mu_{\text{data}}(x) = 0$ for many $x$. This also leads to exploding scores $\nabla \log p_t$ as $t \to 0$ (see Section 5), a behaviour which also observed empirically and further underpins the relevance of the manifold assumption. We will from now on denote the support of $\mu_{\text{data}}$ as *data manifold* $\mathcal{M}$.

A fundamental question is then: What is the support of $\mu_{\text{sample}}$ and how does it compare to $\mathcal{M}$. This is interesting for multiple reasons. On the one hand, if we for example assume that $\mathcal{M}$ is the set of all images of faces, then the knowledge that $\mu_{\text{sample}}$ also has support $\mathcal{M}$ implies that, regardless of how close $\mu_{\text{sample}}$ actually is to $\mu_{\text{data}}$, it will at least always produce an image of a face. On the other hand, we can also compare the support of $\mu_{\text{sample}}$ to that of $\hat{\mu}_{\text{data}}$. The measures $\mu_{\text{sample}}$ and $\mu_{\text{data}}$ sharing their support translates to $\mu_{\text{sample}}$ memorizing the training data and not being able to generalize. Both of these are very valuable insights into the qualities of $\mu_{\text{sample}}$ on its own. Furthermore, statistical distances like the KL-Divergence or the total variation distance are only meaningful if the support of the measures overlap to some degree, as otherwise they will equal the maximum distance value.

If two measures have the same support, they are said to be *equivalent*. Our main result is the following:

**Main result.** *We identify conditions under which $\mu_{sample}$ is able to learn the data manifold, that is $supp(\mu_{sample}) = supp(\mu_{data})$. Applying these results to different settings, we find precise conditions under which an SGM memorizes its training data or under which it is able to learn the right data manifold $\mathcal{M}$.*

*In particular we find that for SGMs to be able to generalize, the approximation error made when approximating the training drift $\nabla \log \hat{p}_t$ has to be unbounded.*

A first illustration of these results is given in Figure 2. We use a simple example, where we can perfectly evaluate the true drift $\nabla \log \pi_t$. We then choose an incorrect initial condition $\mu_{\text{prior}}$ which is far from $p_T$, and also add a constant error to $\nabla \log \pi_t(x)$. The initial measures $\pi_0$ are given as the uniform distribution on the unit sphere and the uniform distribution on 9 samples from the unit sphere in Figure 2a and Figure 2 respectively. We see that the final distribution of the reverse SDE, $\mu_{\text{sample}}$, is not the uniform distribution anymore. This is due to errors in the initial conditions and the drift. Nevertheless, $\mu_{\text{sample}}$ is still supported on the exact same subset as $\pi_0$.

We now proceed as follows. In Section 2 we will summarize some of the most popular forward SDEs that are applied in SGMs. Then, in Section 3 we discuss how the drift approximation $s_\theta(x,t)$ for SGMs is trained in most implementations. We are ready to state our main results in Section 4. In Section 5 we show why and with which speed the true drift $\nabla \log p_t$ of an SGM explodes under the manifold hypothesis. Most of the paper discusses the error in the drift approximation. But as we have discussed, we also have an error in the initial conditions. In Section 6 we will discuss how large the error in the initial conditions will be in practice.

## 2 Popular SDEs used in SGMs

In this section we introduce some of the most popular SDEs that are used when implementing SGMs. 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., 2021b, 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 $\alpha_t$ 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

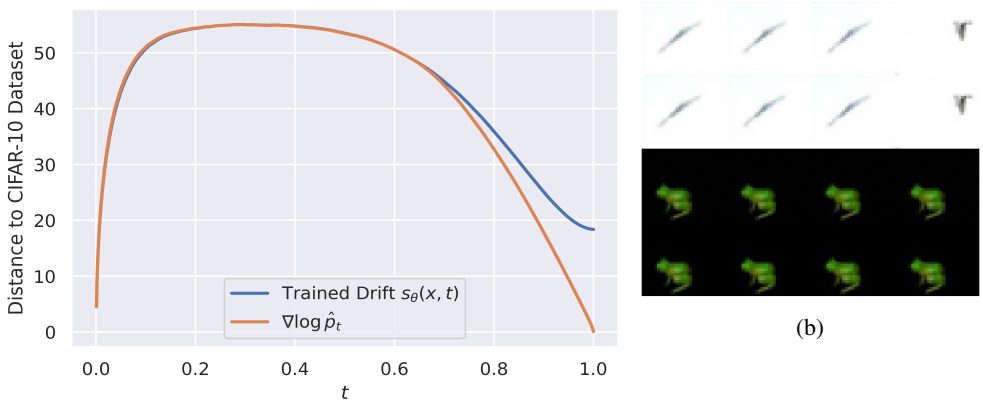
$$\begin{aligned} \mathrm{d}X_t &= V_t, \\ \mathrm{d}V_t &= -X_t - 2V_t + 2\mathrm{d}W_t, \end{aligned}$$

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 $\gamma$. We set both to 1 as they do in their numerical experiments.

## 3 Score approximation with a finite number of samples

We now quickly discuss how the neural network is trained to approximate $\nabla \log p_t$ and what implications this has. The score $\nabla \log p_t$ is approximated by minimizing a variant of

$$L(\theta) = \int_0^T w(t)\, \mathbb{E}_{p_t(x)}[\|\nabla \log p_t(x) - s_\theta(x,t)\|^2]\mathrm{d}t. \tag{3}$$




(a)

(b)

Figure 3: (a): Both lines correspond to the same experiment for different drifts in the reverse SDE. For both lines we started $N = 1000$ paths in the zero vector in $\mathbb{R}^{32\times32\times3}$. For the blue line we used the pretrained CIFAR-10 DDPM++ model from Song et al. [2021b], whereas for the orange line we used the true drift $\nabla \log \hat{p}_t$, which is a mixture of $50000$ Gaussians, one for each training example in CIFAR-10. We then saved the distance from $Y_t$ to the CIFAR-10 training examples, by calculating the distance to the closest example. Above we plot the average distance. We see, that while the reverse SDE run with $\hat{p}_t$ will have a distance of $0$ to the training examples in the end, the SDE with the pretrained drift keeps some distance to the training examples and therefore produces novel images. (b): We evaluate $\nabla \log \hat{p}_t$ as in (a). We do the analogous experiment to Figure 2b on CIFAR-10 and perturb the empirical drift $\nabla \log \hat{p}_t$ with a constant error vector. The first row shows the samples generated by adding the constant error vector $e(x,t) = (1, 1, \dots, 1) \in \mathbb{R}^{32\times32\times3}$ to $\nabla \log \hat{p}_t$. In the second row we searched for the closest image in the CIFAR-10 dataset (with respect to the Euclidean 2-distance on $\mathbb{R}^{32\times32\times3}$) and plotted it. We see that all the sampled images are nearly equal to a corresponding image in CIFAR-10. The distance of the images to their closest image in CIFAR-10 is around $0.07$ for all plotted images. Similar to Figure 2, we can observe the effect of adding the one-vector. The sample distribution $\mu_{\text{sample}}$ got skewed to prefer images that have high pixel values. This corresponds to samples which are mostly white for the human eye. In the third and forth row we repeat the experiment of the first and second row, but add the negative one-vector $e(x,t) = (-1, -1, \dots, -1)$ and get black images.

for some weighting function $w$. This is done via score matching techniques (see [Hyvärinen and Dayan, 2005, Vincent, 2011, Song et al., 2020]). However, the above expectation depends on $p_t$, which we cannot evaluate since it depends on $p_0 = \mu_{\text{data}}$. Nevertheless, we can evaluate the approximation $\hat{p}_t$,

$$\hat{p}_t(x) = \mathbb{E}_{\hat{\mu}_{\text{data}}(x_0)}[p_{t|0}(x|x_0)] = \frac{1}{N}\sum_{i=1}^{N} p_{t|0}(x|x^i) \approx p_t(x) = \mathbb{E}_{\mu_{\text{data}}}[p_{t|0}(x|x_0)]. \tag{4}$$

The surrogate loss

$$\hat{L}(\theta) = \int_0^T w(t)\, \mathbb{E}_{\hat{p}_t(x)}[\|\nabla \log \hat{p}_t(x) - s_\theta(x,t)\|^2]\mathrm{d}t, \tag{5}$$

can be evaluated and used for training. This raises the question in which way or to which degree the loss should actually be minimized. If we would minimize this loss perfectly, then $s_\theta(x,t)$ would be equal to $\nabla \log \hat{p}_t$. The reverse SDE started in an appropriate initial condition with drift $\nabla \log \hat{p}_t$ however, will produce samples from $\hat{\mu}_{\text{data}}$, which are training examples.

## 4 Effects of the approximations

This section contains our main results. We first state the assumptions we have to make and then the Theorems.

### 4.1 Error in the initial condition

The following assumption is needed for the reverse SDE to be defined even in the case when the initial distribution $\pi_0$ has a degenerate support. In Lemma 1 we will show that all the SDEs from Section 2 satisfy the Assumption.

**Assumption 1.** *There is a constant $C$ such that*

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

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

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

*Furthermore, for each $S \in (0, T)$ and all $x, y$ for which $\|x\|, \|y\| \leq N$ there is a constant $C_{S,N}$ such that*

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

Conditions $(i)$-$(iii)$ are technical conditions on the forward SDE. They ensure that if we run a solution to the forward SDE, $X_t$, backwards in time, then $X_t^R := X_{T-t}$ 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). The following result shows that Assumption 1 can be expected to hold in practice. To simplify the calculations we assume that the data manifold $\mathcal{M} = \mathrm{supp}(\mu_{\text{data}})$ 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.

**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 fulfil Assumption 1.*

We are now ready to state our first main result,

**Theorem 1.** *Denote by $\pi_t$ the marginals of the forward SDE started in $\pi_0$. Assume that Assumption 1 holds and that $\mu_{prior}$ is absolutely continuous with respect to $\pi_T$. 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 as $\mu_{sample}$. The distribution $\mu_{sample}$ is absolutely continuous with respect to $\mu_{data}$. If $\pi_T$ and $\mu_{prior}$ are equivalent, then so are $\mu_{sample}$ and $\pi_0$.*

*(ii)* *Furthermore, for any $f$-divergence $D_f$,*

$$D_f(\mu_{sample}|\pi_0) \leq D_f(\mu_{prior}|\pi_T) \quad \textit{and} \quad D_f(\pi_0|\mu_{sample}) \leq D_f(\pi_T|\mu_{prior}).$$

Applying the above theorem with $\pi_T = p_T$ tells us something about the equivalence and distance between $\mu_{\text{sample}}$ and $\mu_{\text{data}}$, since $\pi_0 = p_0 = \mu_{\text{data}}$. Applying the theorem with $\hat{p}_t$ tells us something about the generalization capabilities of SGMs.

The measures $\pi_T$ and $\mu_{\text{prior}}$ are normally both supported on all of $\mathbb{R}^d$ and therefore equivalent. Since Assumption 1 also holds in most situations (see Lemma 1), the requirements for Theorem 1 are satisfied in practice. From item $(i)$ with $\pi_t = \hat{p}_t$ we can conclude that the error we make in the initial conditions is not responsible for the generalization capacities of SGMs.

The second item then shows that the $f$-divergences between $\mu_{\text{sample}}$ and $\mu_{\text{data}}$ are bounded by the $f$-divergences of $\mu_{\text{prior}}$ to $p_T$. The total variation distance and the KL-Divergence are both special cases of $f$-divergences.

### 4.2 Error in the drift

Given a forward SDE with marginals $\pi_t$ and an approximation $s(x, t)$ to $\nabla \log \pi_t$, we define the reverse SDE for the approximation $s$ as

$$\begin{aligned} \mathrm{d}\tilde{Y}_t &= -\beta(\tilde{Y}_t)\mathrm{d}t + \sigma\sigma^T s(\tilde{Y}_t, t)\mathrm{d}t + \sigma \mathrm{d}B_t, \\ \tilde{Y}_0 &\sim q_0. \end{aligned} \tag{6}$$

**Assumption 2.** *We assume that the reverse SDE $\tilde{Y}_t$ has a solution on $[0, T)$. For $t < T$, we define the Girsanov weights*

$$Z_t = \exp\left(\int_0^t \sigma^T(s(\tilde{Y}_t, t) - \nabla \log \pi_t(\tilde{Y}_t)) \cdot \mathrm{d}B_s - \frac{1}{2}\int_0^t \|\sigma^T(s(\tilde{Y}_t, t) - \nabla \log \pi_t(\tilde{Y}_t))\|^2 \mathrm{d}s\right) \tag{7}$$

*and assume that the $Z_t$ are a uniformly integrable martingale.*

We shortly discuss this assumption. The assumption that $Z_t$ is a martingale is equivalent to the expectation of $Z_t$ being equal to 1 for all $t$. The assumption that it is uniformly integrable is more technical (see Appendix A.2), but is fulfilled for example if for some $p > 1$, $\mathbb{E}[|Z_t|^p] < \infty$ for all $t$. For example, if the $Z_t$ have bounded variance, then they are uniformly integrable.

A condition that ensures that $Z_t$ is both, a martingale and uniformly integrable, is given by *Novikov's condition* Novikov [1980]. It states, that if

$$N = \mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^T \|\sigma^T(s(x, t) - \nabla \log \pi_t)\|^2 \mathrm{d}s\right)\right] < \infty, \tag{8}$$

then $Z_t$ is a uniformly integrable martingale.

Using Assumption 2 we can now state

**Theorem 2.** *Assume that Assumption 2 holds. Assume furthermore that Assumption 1 holds with $\nabla \log \pi_t$ replaced by $s(x, t)$.*

*Then $\tilde{Y}_T = \lim_{t \to T} \tilde{Y}_t$ is well defined. Moreover, its distribution is equivalent to the distribution of $Y_T$. In particular, if $\|s(x, t) - \nabla \log \hat{p}_t\|$ is bounded, then Assumption 2 holds, and the SGM has memorized its training data.*

Putting Theorem 1 and Theorem 2 together, we can conclude that, if both Assumptions hold, $\mu_{\text{sample}}$ is equivalent to $\pi_0$. Therefore, if Assumption 2 holds for $\pi_t = p_t$, we have a positive statement and know that $\mu_{\text{sample}}$ will have the exact same support as $\mu_{\text{data}}$, i.e. it has learned the data manifold $\mathcal{M}$.

If the Assumption would hold for $\pi_t = \hat{p}_t$, we would however just memorize the training data, see Figure 2b or Figure 3b for a visualization. However, empirically it has been shown that SGMs are able to create novel samples (see, Figure 3a or Dhariwal and Nichol [2021]). Therefore, we can deduce that Assumption 2 is violated in practice.

SGMs are trained by minimizing $\hat{L}(\theta)$ (see (5)). On the other hand, if $L(\theta)$ is minimized well enough for $Z_t$ to fulfil Assumption 2, then the SGM is collapsed onto its training data. This is especially interesting since

$$L(\theta) = \mathbb{E}_{Y_0 \sim \hat{p}_0}[\log(Z_T)].$$

We evaluated N from (8) on CIFAR-10, once for the difference between the $s_\theta(x, t)$ from Song et al. [2021b] and $\nabla \log \hat{p}_t$, and once by just using a perturbed drift with a constant error, $s(x, t) = \nabla \log \hat{p}_t + \frac{1}{2}(1, 1, \ldots, 1)$, see Figure 4. The reverse SDE using the drift $\nabla \log \hat{p}_t + \frac{1}{2}(1, 1, \ldots, 1)$, is equivalent to $\hat{\mu}_{\text{data}}$, as we know from Theorem 2 and have also already observed in Figure 3b. Figure 4 confirms that $Z_t$ is indeed a uniformly integrable martingale and therefore fulfils Assumption 2.

In future work we believe that further understanding the characteristics of $Z_t$, how they relate to the minimization of $L(\theta)$ and generalization is crucial to the understanding of SGMs and their empirical success. Recent works also study the question on how the neural network architecture and parametrization is related to the boundedness of the output Kim et al. [2022]. The relationship between these choices and the properties of $Z_t$ is also an interesting research avenue. Lastly, the distribution of $Z_t$ is very heavy-tailed. Most of the samples are very small with a few extremely large samples in between. Therefore one needs many samples from $Z_t$ to understand its characteristics. Finding robust estimators for $Z_t$ or its expectation could help their usage in the training or evaluation procedure.

## 5 Drift Explosion under manifold hypothesis

First, we note that all SDEs in Section 2 are linear SDEs. Therefore, their transition kernels are Gaussian ([Pavliotis, 2014, Section 3.7]):

$$\pi_t(X_t = x | X_0 = x_0) = \mathcal{N}(x; m_t(x_0), \Sigma_t).$$

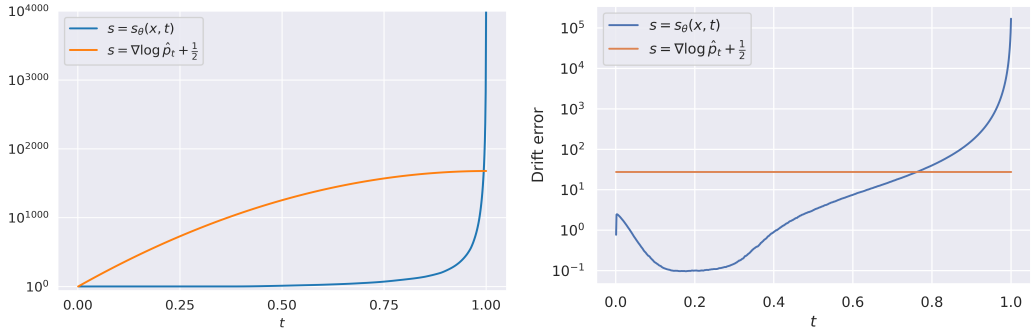



Figure 4: Left: We simulated the reverse SDE on CIFAR-10, once with the pretrained CIFAR-10 DDPM++ model $s_\theta$ from Song et al. [2021b] and once with a perturbed drift $s(x,t) = \nabla \log \hat{p}_t + \frac{1}{2}(1,1,\dots,1)$. We then evaluated the integral (8) numerically. For the perturbed drift, the integral does not seem to explode as $t \to 1$, implying that $Z_t$ is a martingale. We see that for the DDPM++ drift, the integral explodes, therefore we can not infer that $Z_t$ is a martingale. We used $N = 12000$ simulations from both of the SDEs to generate this plot. Right: We again ran the two SDEs with the drifts as in the left Figure. This time, we measured the distance to the empirical drift $\|s(x,t) - \nabla \log \hat{p}_t(x)\|$. We repeated the experiment $N = 2560$ times and plotted the mean distance. For the constant perturbation we also of course get a constant distance. The distance of the true drift to $\nabla \log \hat{p}_t$ is initially very small but explodes as $t \to 1$. From our results we know that this explosion is necessary for the SGM to generalize.

The explicit form of $m_t$ and $\Sigma_t$ differ for each of the SDEs and can be found in Appendix C.1. We remark that $\Sigma_t$ does not depend on the initial condition $x_0$. The transition kernel above is the distribution of the SDE started in a single point $x_0$. Since we start the SDE in $\mu_{\text{data}}$ we need to average over $\mu_{\text{data}}$ to get the marginal at time $t$:

$$p_t(x) = \int_{\mathbb{R}^d} \mathcal{N}(x; m_t(x_0), \Sigma_t)\mu_{\text{data}}(x_0)\mathrm{d}x_0. \tag{9}$$

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

$$\nabla \log p_t(x) = \frac{\nabla p_t(x)}{p_t(x)} = \Sigma_t^{-1}(x - \mathbb{E}[m_t(X_0)|X_t = x]). \tag{10}$$

We now want to evaluate $\nabla \log p_t$ along a typical path of $Y_t$, i.e. we are interested in $\mathbb{E}[\nabla \log p_t(Y_t)]$. The distribution of $Y_t$ however depends on the drift approximation $s_\theta$ we use. For this calculation we will assume that we are able to run the reverse SDE with the true drift $s_\theta(x,t) = p_t(x)$. Then however, since $Y_t$ is then just $X_t$ run backwards, they have the same distributions and we can calculate

$$\mathbb{E}[\|\nabla \log p_t(Y_t)\|] = \mathbb{E}[\|\nabla \log p_t(X_t)\|] = \Sigma_t^{-1}\mathbb{E}[\|X_t - \mathbb{E}[m_t(X_0)|X_t = y]\|].$$

If the manifold $\mathcal{M}$ is not to badly behaved, we can expect that for small $t$ and almost all $x$, $\mathbb{E}[X_0|X_t = x]$ to be very close to the data manifold $\mathcal{M}$. Especially, $\|X_t - \mathbb{E}[X_0|X_t = x]\|$ will be larger than $\text{dist}(X_t, \mathcal{M})$ in that case. However, the distribution of $X_t$ can be represented as $m_t(X_0) + \sqrt{\Sigma_t}\xi$, where $\xi$ has a standard normal distribution. For the SDEs we treated in Section 2, $m_t(X_t)$ is either equal or very close to $m_t(X_t) = X_t$ for small values of $t$. Finally, if we assume that $\mathcal{M}$ is a subset of relatively low dimension in a high dimensional space, we can expect that with very high probability $\sqrt{\Sigma_t}\xi$ points away from the data manifold. Therefore the distance of $X_t$ to $\mathcal{M}$ can be approximated by $\sqrt{\Sigma_t}\xi$. Putting these approximations together we can calculate

$$\mathbb{E}[\|X_t - \mathbb{E}[m_t(X_0)|X_t = y]\|] \gtrsim \mathbb{E}[\text{dist}(X_t, \mathcal{M})] \gtrsim \|\Sigma_t\|^{1/2}\mathbb{E}[\|\xi\|] \approx \|\Sigma_t\|^{1/2}\sqrt{d},$$

where $d$ is the dimension of the data space in which the samples $x_i$ lie. Therefore we can conclude that

$$\|\nabla \log p_t(Y_t)\| \gtrsim \frac{d^{1/2}}{\|\Sigma_t\|^{1/2}}.$$

For the Brownian motion for example, $\Sigma_t = t$ and therefore the right hand side scales like $\frac{1}{\sqrt{t}}$. If $\Sigma_t$ is the covariance of $p_t$, then we can expect $\nabla \log p_t(Y_t)$ to be of order $\frac{1}{\|\Sigma_t\|^{1/2}}$. Furthermore,

$\nabla \log p_t(Y_t)$ will point towards the data manifold for small $t$. The drift $\nabla \log p_t(Y_t)$ then acts like a *support matching force*, where the force grows to infinity as $t \to 0$, absorbing all the SDE paths onto the manifold.

# 6 Distance from $p_T$ to $\mu_{\text{prior}}$

We have seen in Theorem 1 that the distance between $\mu_{\text{sample}}$ and $\mu_{\text{data}}$ is directly related to the distance between $p_T$ and $\mu_{\text{prior}}$ if we neglect the errors made in the approximation of the drift. For both, the OU-Process and the CLD, there are plenty of results on the distance of $p_t$ to $\mathcal{N}(0, I_d)$. In general, one can expect this distance to grown exponentially in time $t$.

The Brownian motion however does not converge to a stationary distribution and therefore one has to choose a different $\mu^T_{\text{prior}}$ for each $T$ to approximate $p_T$. In practice, one normally chooses a normal distribution $p_T = \mathcal{N}(m_t, C_t)$ (see [Song et al., 2021b, Appendix C]). The following Lemma is derives the optimal values for $m_t$ and $C_t$.

**Lemma 2.** *Let $p_T$ be the $T$-time marginal of the Brownian motion process $X_t$ of Section 2. 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 $C_T = \mathrm{Cov}[\mu_{data}] + TI_d$. If we restrict the covariance to be a multiple of the identity matrix, 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 moments. We prove it in Appendix E.2. The following result shows that the distance between $p_T$ and $\mu^T_{\text{prior}}$ decreases with time and also gives a rate. It justifies using the Brownian motion and a normal prior distribution for SGMs.

**Lemma 3.** *Let $p_t$ be the time $t$-marginal of a Brownian motion with initial condition $\mu_{data}$. Denote by $c_i, i = 1, \dots, d$, the eigenvalues of $\mathrm{Cov}(\mu_{data})$. Let $\mu^T_{prior}$ be the normal distribution with mean $m_T = \mathbb{E}[\mu_{data}]$ and covariance $C_T = \mathrm{Cov}[\mu_{data}] + TI_d$. Then*

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

The proof can be found in Appendix E.2.

# 7 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.

# 8 Conclusion

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

Since one can not actually access the full support of $\mu_{\text{data}}$, 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


G. Batzolis, J. Stanczuk, C. Schönlieb, and C. Etmann. Conditional image generation with score-based diffusion models. *CoRR*, abs/2111.13606, 2021. URL https://arxiv.org/abs/2111.13606.

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.

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. Kim, S. Shin, K. Song, W. Kang, and I. Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 11201–11228. PMLR, 2022. URL https://proceedings.mlr.press/v162/kim22i.html.

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.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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.

A. A. Novikov. On conditions for uniform integrability of continuous non-negative martingales. *Theory of Probability & Its Applications*, 24(4):820–824, 1980.

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

S. Peluchetti. Non-denoising forward-time diffusions. 2021.

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, C. Durkan, I. Murray, and S. Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021a.

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, 2021b. 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 7.
   (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]

# A Stochastic prerequisites

In this section we give a formal introduction to some of the concepts used in this work. For a more rigorous treatment, see for example [Klenke, 2013] or [Karatzas and Shreve, 2012].

## A.1 Equivalence of measures / Girsanov Theorem

First we define absolute continuity of measures. Let $\mu$ and $\nu$ be two measures on $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a $\sigma$-algebra.

**Definition 1.** *We say that $\mu$ is absolutely continuous with respect to $\nu$ if $\mu(A) = 0$ for any $A \in \mathcal{F}$ such that $\nu(A) = 0$. We also denote this by $\mu \ll \nu$.*

Two measures $\mu$ and $\nu$ are equivalent if $\mu \ll \nu$ and $\nu \ll \mu$. Loosely speaking, we can say that $\mu \ll \nu$ if the support of $\mu$ is contained in the support of $\nu$ and they are equivalent if they share the same support.

The Radon-Nikodym theorem tells us that if $\mu \ll \nu$, then under mild conditions there exists a density $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} : \Omega \to \mathbb{R}$ such that $\mu(A) = \int_A \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\omega)\mathrm{d}\nu(\omega)$. Therefore, we can obtain $\mu$ through a reweighting of $\nu$. One specific instance of this is the Girsanov Theorem. Assume we are given the solutions to two SDEs in $\mathbb{R}^d$,

$$\mathrm{d}Y_t = b(t, Y_t)\mathrm{d}t + \sigma(t, Y_t)\mathrm{d}W_t \tag{11}$$

and

$$\mathrm{d}\tilde{Y}_t = b(t, \tilde{Y}_t)\mathrm{d}t + \sigma(t, \tilde{Y}_t)e(t, \tilde{Y}_t)\mathrm{d}t + \sigma(t, \tilde{Y}_t)\mathrm{d}B_t. \tag{12}$$

Both of these induce a measure on the space of continuous functions $\Omega = C([0, T], \mathbb{R}^d)$. We denote them by $\mathbb{P}$ and $\tilde{\mathbb{P}}$ respectively. Then the Girsanov Theorem equips us with conditions under which the measures $\mathbb{P}$ and $\tilde{\mathbb{P}}$ are equivalent. Furthermore, in case of equivalence we get a formula for the density of $\tilde{\mathbb{P}}$ with respect to $\mathbb{P}$. The relative density is given as

$$Z_T = \exp\left(\int_0^T e(s, Y_s)\mathrm{d}W_s - \frac{1}{2}\int_0^T \|e(s, Y_s)\|^2 \mathrm{d}s\right).$$

For a full statement of the Girsanov Theorem and under which conditions it holds, see [Karatzas and Shreve, 2012, Section 3.5].

## A.2 Uniform integrability

Since we are treating the case where the drift explodes as $t \to T$ we end up with densities

$$Z_t = \exp\left(\int_0^t e(s, Y_s)\mathrm{d}W_s - \frac{1}{2}\int_0^t \|e(s, Y_s)\|^2 \mathrm{d}s\right). \tag{13}$$

on $C([0, t], \mathbb{R}^d)$, but not with a density on $C([0, T], \mathbb{R}^d)$. Uniform integrability is exactly the condition one needs to extend these local densities.

**Definition 2.** *A family $\{X_\alpha\}$ of random variables is called uniformly integrable if*

$$\sup_\alpha \mathbb{E}\left[|X_\alpha| \, 1_{\{|X_\alpha| > s\}}\right] \to 0$$

*as $s \to \infty$.*

In the proof of Theorem 2 we implicitly use the following two results which we here state as a lemma. The filtration $\mathcal{F}_t$ is defined as in the proof of Theorem 2.

**Lemma 4.** *Assume the $Z_t$ in* (13) *form a uniformly integrable martingale on $[0, T)$. Then,*

- *the limit $\lim_{t \to T} Z_t$ exists in $L^1$. We denote this limit by $Z$.*
- *Furthermore, $\tilde{\mathbb{P}}$ is absolutely continuous with respect to $\mathbb{P}$ on $\mathcal{F} = \sigma(\cup_{t<T}\mathcal{F}_t)$ with density $Z$.*

*Proof.* Both of these results are standard. The first one can for example be found in [Karatzas and Shreve, 2012, Section 1.3.B]. For the second one we compute that for any $A \in \mathcal{F}_s$,

$$\mathbb{E}_{\mathbb{P}}[1_A Z] = \mathbb{E}_{\mathbb{P}}[1_A \lim_{t\to T} Z_t] = \lim_{t\to T} \mathbb{E}_{\mathbb{P}}[1_A Z_t] = \mathbb{E}_{\mathbb{P}}[1_A Z_s] = \mathbb{E}_{\tilde{\mathbb{P}}}[1_A],$$

where we used $L^1$ convergence in the second equality and the martingale property of $Z_s$ in the third equality. Therefore $Z$ is a density of $\tilde{\mathbb{P}}$ with respect to $\mathbb{P}$ on each $\mathcal{F}_s$ for $s < T$. Therefore $Z$ is also a density of $\tilde{\mathbb{P}}$ with respect to $\mathbb{P}$ on $\mathcal{F}$ which concludes the proof. □

# B Numerics

All numerical experiments can be run on a consumer grade computer within a few minutes.

## B.1 Figure 1

We first discuss the top left figure of Figure 1. We set $p_0 = \mu_{\text{data}}$ to a mixture of two Gaussian $\mathcal{N}(-2, \frac{1}{100})$ and $\mathcal{N}(2, \frac{1}{100})$ with weights $w_1 = \frac{1}{3}$ and $w_2 = \frac{2}{3}$ respectively. Then, we draw $N = 5\,000\,000$ samples from $\mu_{\text{data}}$, denoted by $Y_0^n$, $n = 1, \dots, N$. An Euler-Maruyama discretization of the Brownian motion propagates these samples from time $t = 0$ to $t = 1$ by

$$X_{i+1}^n = X_i^n + \sqrt{dt} Z_i^n,$$

where $Z_n^i \sim \mathcal{N}(0,1)$ are i.i.d. random variables, independent of $X_m^j$ for $m \le n$ and $j = 1, \dots, N$. The time index $i$ runs from 0 to $I = 2000$ and $dt$ is set to $dt = \frac{1}{I}$. The initial samples $\{X_0^n\}_{n=1}^N$ are used to create the left line plot of $p_0$ and the final samples $\{X_I^n\}_{n=1}^N$ are used to create the right line plot of $p_1$ using kernel density estimation. The $\{X_i^n\}_{n=1}^N$ are approximate samples from $p_{i/I}$. Therefore, we create histograms using $\{X_i^n\}_{n=1}^N$ to approximate $p_{i/I}$. The height of the histogram bars corresponds to the square root of the colour intensity in the heat map. The horizontal axis in the heat map stands for the time $t$, whereas the vertical axis stands for the position $x$. At location $(t, x)$ we plot an estimate of $\sqrt{p_t(x)}$. We apply the square root since it improves the contrast in areas where $p_t(x)$ is close to 0 and makes it more visible where $p_t(x) > 0$ to the observer.

For the bottom left figure we show the same plots, just for the reverse SDE (6) instead of the forward SDE. Since the initial distribution is a Gaussian mixture we can exactly calculate $p_t$ using

$$p_t(x) = w_1 \mathcal{N}(x; m_1, s_1^2 + t) + w_2 \mathcal{N}(x; m_2, s_2^2 + t), \tag{14}$$

where we use $\mathcal{N}(x; m, v)$ for the probability density function of a normal distribution with mean $m$ and variance $v$, evaluated at $x$. With the above expression of $p_t$ one could compute an analytical representation of $\nabla \log p_t$. We use automatic differentiation instead. The reverse SDE (6) is simulated with a disturbance $e(x,t) = 1$ and initial condition $q_0 = \mu_{\text{prior}} = \mathcal{N}(0,1)$. The Euler-Maruyama method is run with the same step size $dt = \frac{1}{I}$. More precisely, the one step transition kernel of the discretized reverse SDE is

$$Y_{i+1}^n = Y_i^n + dt\, (\nabla \log p_{1-\frac{i}{I}}(Y_i^n) + 1) + \sqrt{dt}\, \tilde{Z}_i^n, \tag{15}$$

where $\tilde{Z}_n^i \sim \mathcal{N}(0,1)$ are i.i.d. random variables, independent of $Y_m^j$ for $m \le n$ and $j = 1, \dots, N$. The plots are created in the same way as for the upper left plot, except that we reverse the time axis to plot $p_t$ and $q_{1-t}$ directly underneath each other.

On the right side we plot the same kernel density estimates already plotted on the left side as $\mu_{\text{sample}}$ and $\mu_{\text{data}}$ into the same plot for comparison.

## B.2 Figure 2b and 2a

Figure 2b is created by setting $\mu_{\text{data}}$ to be the uniform distribution on $M = 9$ equally spaced samples $\{x_i\}_{i=1}^M$ on the unit sphere $\mathcal{S}^1$. This can also be viewed as a Gaussian mixture with 9 components where each component having mean $x_i$ and variance 0. Therefore, we can again explicitly calculate $p_t$ for $t > 0$ as in (14),

$$p_t(y) = \frac{1}{M} \sum_{i=1}^M \mathcal{N}(y; x_i, t).$$

| | $m_t(z_0)$ | $\Sigma_t$ |
|---|---|---|
| Brownian Motion | $z_0$ | $tI_d$ |
| OU-Process | $\exp(-t)z_0$ | $(1-\exp(-2t))I_d$ |

Table 1: The mean and covariance of the Gaussian transition kernels of the Brownian Motion and the Ornstein-Uhlenbeck process SDEs.

The score $\nabla \log p_t(y)$ is evaluated using automatic differentiation. The reverse SDE 6 is simulated with $q_0 = \mathcal{N}((x = -1.5, y = 0), I_2)$ and $e((x, y), t) = (x = 0, y = -1)$. For the numerical simulation, we again use the Euler-Maruyama scheme. We use a step width of $dt = \frac{0.9}{1000} \approx \frac{1}{1000}$ for $t = [0, 0.9]$, $dt = \frac{0.09}{1000} \approx \frac{1}{10\,000}$ for $t \in [0.9, 0.99]$ and of $dt = \frac{1}{100\,000}$ for $t \in [0.99, 1]$. For a simulation of $N = 50\,000$ paths of the reverse SDE, we start by drawing $Y_0^n = A_n + Z_n$, where $A_n$ are i.i.d. uniformly distributed on $\{x_i\}_{i=1}^M$ and $Z_n \sim \mathcal{N}(0, 1)$ i.i.d.. The $\{A_n\}$ and $\{Z_n\}$ are also independent from each other. We then propagate the $Y_0^n$ similarly to (15), except that we use a different values for $dt$ depending on $t$. This leads to approximate samples $Y_t^i$ from $q_t$. At the displayed times $t$ we plot the function

$$h_t(x) = \sum_{i=1}^{N} k(x, Y_t^i),$$

where $k$ is an unnormalized Gaussian kernel with a very small bandwidth parameter,

$$k(x, y) = \exp(-1000\|x - y\|^2).$$

Normalizing $h_t$ gives us a density estimate of $q_t$. We plot these estimates as heat maps for different values of $t$.

For Figure 2a we follow the same steps as for Figure 2b, except that $\mu_{\text{data}}$ is set to the uniform distribution over $M = 256$ evenly spaced samples from the unit sphere $\mathcal{S}^1$.

### B.3 Figure 3

For Figure 3 we used the DDPM++ model from the Github repository for the paper from Song et al. [2021b]. Then we evaluated the true score using (4), where the sum runs through the $N = 50000$ training examples of CIFAR-10, Krizhevsky et al. [2009]. A similar experiment using the true marginals $\hat{\mu}_{\text{data}}$ has also been conducted in Peluchetti [2021].

## C Studying the forward Densities

### C.1 Transition kernels

The transition kernels $p(z_0, \cdot)$ for the SDEs from Section 2 are of the from

$$q_0(z_0, \cdot) = \mathcal{N}(m_t(z_0), \Sigma_t),$$

where $m_t$ and $\Sigma_t$ are given in Table 1 for the Brownian Motion and the Ornstein-Uhlenbeck process. The form of the transition kernels for the CLD are more involved. They can be found in Dockhorn et al. [2021, Appendix B.1].

**Lemma 5.** *The marginal densities $p_t(x)$ of the SDEs treated in Section 2 depend smoothly on $x$ and $t$.*

*Proof.* One can combine the form of $m_t$ and $\Sigma_t$ and the explicit representation of $p_t$ in (9) to see this. More generally, for the Brownian motion and the OU-Process this is a result of the Hörmander theorem. For the CLD it is a result of hypocoercivity. □

### C.2 Form of the drift

We now prove that we can represent the drift as in (10).

**Lemma 6.** *Assume that $p_t$ has the form* (9)*, i.e.*

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

*Then* (10) *holds true, i.e.*

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

*Proof.*

$$\begin{aligned}
&\nabla p_t(z) \\
&= \frac{1}{\sqrt{\det(2\pi\Sigma_t)}} \int_{\mathbb{R}^d} \nabla \exp\left(-\frac{1}{2}(z - m_t(z_0))\Sigma_t^{-1}(z - m_t(z_0))\right)\mu_{\text{data}}(z_0)\mathrm{d}z_0 \\
&= \frac{1}{\sqrt{\det(2\pi\Sigma_t)}} \int_{\mathbb{R}^d} \Sigma_t^{-1}(z - m_t(z_0)) \exp\left(-\frac{1}{2}(z - m_t(z_0))\Sigma_t^{-1}(z - m_t(z_0))\right)\mu_{\text{data}}(z_0)\mathrm{d}z_0 \\
&= \Sigma_t^{-1} z p_t(z) \\
&\quad - \frac{1}{\sqrt{\det(2\pi\Sigma_t)}} \int_{\mathbb{R}^d} \Sigma_t^{-1} m_t(z_0) \exp\left(-\frac{1}{2}(z - m_t(z_0))\Sigma_t^{-1}(z - m_t(z_0))\right)\mu_{\text{data}}(z_0)\mathrm{d}z_0.
\end{aligned}$$

If we now divide everything by $p_t$ it cancels in the first summand. In the second summand we get the formula for the conditional expectation (see, for example [Klenke, 2013, Section 8.2]). □

# D Reverse SDEs: The general case

One can also treat more general forward SDEs than we did in Section 1. This leads to a more complicated form of the reverse SDE. Our Theorems do not use the specific form of the forward SDEs and therefore also hold in the general case. We denote the forward SDE by

$$\begin{aligned}
\mathrm{d}X_t &= \beta(t, X_t)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}W_t, \\
X_0 &\sim \mu_{\text{data}},
\end{aligned} \tag{16}$$

where $\mu_{\text{data}}$ is supported on $\mathcal{M} \subset \mathbb{R}^d$ and $W_t$ is a $\mathbb{R}^r$ valued Brownian motion. The drift $b$ maps from $\mathbb{R} \times \mathbb{R}^d$ to $\mathbb{R}^d$. The dispersion coefficient $\sigma$ maps from $\mathbb{R} \times \mathbb{R}^d$ to the $d \times r$-matrices. The time-reversed process $Y_t := X_{T-t}$ is then a solution to

$$\begin{aligned}
\mathrm{dY}_t &= b(t, Y_t)\mathrm{d}t + \sigma(T - t, Y_t)\mathrm{d}B_t, \\
Y_0 &\sim q_0,
\end{aligned} \tag{17}$$

with

$$b_i(t, y) = -\beta(T - t, y) + \frac{\sum_j \nabla_j(a_{\text{ij}}(T - t, y)p_{T-t}(y))}{p_{T-t}(y)}, \quad a(t, y) = \sigma(t, y)\sigma(t, y)^T,$$

and $q_0 = p_T$, see Haussmann and Pardoux [1986]. This simplifies to the case discussed in Section 1 if $r = d$, $\sigma$ is is a multiple of the identity matrix and $\beta$ is independent of the time $t$. In Assumption 1 we treat the case where the SDEs are of the form written-out in Section 1. The items $(i) - (iii)$ need to be replaced by their more general counterpart as found in Haussmann and Pardoux [1986, Section 2]. In the last item $(iv)$, $\nabla \log p_t$ needs to be replaced by $\frac{\sum_j \nabla_j(a_{\text{ij}}(T-t,y)p_{T-t}(y))}{p_{T-t}(y)}$.

# E Proofs

### E.1 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 $\mu_{\text{data}}$ is degenerate and supported on a low-dimensional

substructure $\mathcal{M}$. As we have seen in Section **??**, this can lead to an exploding drift in the reverse SDE as $t \to T$. Nevertheless, in order to understand the properties of $\mu_{\text{data}}$, 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.

### E.1.1 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) started in $p_0 = \mu_{\text{data}}$. $P$ has marginals $p_t$. Denote by $X_t$ the canonical projections $X_t(\omega) = \omega(t)$ for $\omega \in \Omega$. We define $Q$ through

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

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

$$KL(q_T|\mu_{\text{data}}) \leq KL(Q|P) = KL(\mu_{\text{prior}}|\pi_T). \tag{18}$$

It remains to prove that by running $Q$ backwards we obtain a solution to (2) started in $\mu_{\text{prior}}$. 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 $\mu_{\text{prior}}$ 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 (18). □

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 $\Omega$, 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 (17) inherits these properties.

### E.1.2 Theorem 2

*Proof.* Denote the space $C([0,T),\mathbb{R}^d)$ by $\Omega$. We also define the natural filtration $\mathcal{F}_t = \sigma(x(s)|s \leq t)$. We denote the distribution of $Y$ on $\Omega$ 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 (6) 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 (6), 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 (6) 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.

## E.2 Proof of the Lemmas

We start by proving Lemma 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 $C^\infty$ 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 $(iv)$. □

We now prove Lemma 2.

*Proof.* We have that $X_0 \sim \mu_{\text{data}}$. Denote the mean and covariance of $\mu_{\text{data}}$ by $a$ and $C$ respectively. We define

$$n_t = \mathcal{N}(m_t, V_t)$$

for some functions $m_t$ and $V_t$. If $V_t$ would not have full rank, $n_t$ would be a degenerate distribution. Since $p_t > 0$ almost everywhere for $t>0$, the KL divergence from $p_t$ to $n_t$ would be infinite. We can therefore restrict $V_t$ to be an invertible matrix. We denote the entropy of $p$ by $H$, $H(p) = -\int \log(p)p\mathrm{d}x$.

We can write $X_t$ as $X_t = X_0 + \sqrt{t}Z$ where $Z \sim \mathcal{N}(0, I_d)$. Now for the KL-divergence at time $t$ it holds that

$$\begin{aligned} KL(p_t|n_t) &= -H(p_t) - \int \log(n_t(x))p_t(x)\mathrm{d}x \\ &= -H(p_t) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log\left(\det(V_t)\right) + \frac{1}{2}\int (x-m_t)^T V_t^{-1}(x-m_t)\, p_t(x)\mathrm{d}x. \end{aligned}$$

We compute

$$\begin{aligned} &\int (x-m_t)^T V_t^{-1}(x-m_t)\, p_t(x)\mathrm{d}x \\ &= \mathbb{E}[(X_t-m_t)^T V_t^{-1}(X_t-m_t)] = \mathbb{E}[(X_0+\sqrt{t}Z-m_t)^T V_t^{-1}(X_0+\sqrt{t}Z-m_t)] \\ &= \mathbb{E}[(X_0-a)^T V_t^{-1}(X_0-a)] + (a-m_t)^T V_t^{-1}(a-m_t) + t\mathbb{E}[Z^T V_t^{-1} Z] \\ &= \mathrm{tr}(CV_t^{-1}) + t\,\mathrm{tr}(I_d V_t^{-1}) + (a-m_t)^T V_t^{-1}(a-m_t). \end{aligned}$$

We added and subtracted $a$ in the second equality. In the last equality we used that $\mathbb{E}[X^T A X] = \mathrm{tr}(\mathrm{Cov}(X)A)$ if $X$ is a centred random variable. Substituting this into our prior calculation, we arrive at

$$\begin{aligned} &KL(p_t|n_t) \\ =\ & -H(p_t) + \frac{1}{2}\log\left(\det(2\pi V_t)\right) + \frac{1}{2}\left(\mathrm{tr}((C+tI_d)V_t^{-1}) + (a-m_t)^T V_t^{-1}(a-m_t)\right). \end{aligned}$$

The mean function $m_t$ only appears in the last term. Therefore, we decrease the KL-divergence by setting $m_t = a$. We now optimize for the covariance matrix $V_t$. The $H(p_t)$ term is independent of $V_t$. Consequently, to optimize the $KL(p_t|n_t)$ with respect to $V_t$, it is enough to optimize

$$\begin{aligned} L(V_t^{-1}) &= \log\left(\det(V_t)\right) + \mathrm{tr}((C+tI_d)V_t^{-1}) \\ &= -\log\left(\det(V_t^{-1})\right) + \mathrm{tr}((C+tI_d)V_t^{-1}), \end{aligned}$$

where we used that $\det(A^{-1}) = \det(A)^{-1}$ to rewrite the loss in terms of $V_t^{-1}$. Since $-\log(\det(\cdot))$ is convex on the positive semidefinite matrices and the other summands are linear, $L$ is convex in $V_t^{-1}$. We take the gradient of $L$ with respect to the Frobenius inner product and obtain

$$\nabla_{V_t^{-1}} L(V_t^{-1}) = -V_t + C + tI_d.$$

Setting the gradient to 0 has the unique solution

$$V_t = C + tI_d,$$

which is a local and global minimum due to the convexity of $L$.

If we restrict $V_t$ to be of the form $V_t = v_t I_d$, the loss becomes

$$L(v_t^{-1}) = -d\log(v_t^{-1}) + v_t^{-1}\mathrm{tr}(C) + v_t^{-1}td.$$

By setting the derivative of this to zero, we arrive at

$$v_t = \mathrm{tr}(C)/d + t.$$

□

We now proceed to prove Lemma 3. For that we first need the following proposition.

**Proposition 1.** *Let $p_t$ be the time $t$-marginal of a Brownian motion started in $\mu_{data}$. Assume that the support of $\mu_{data}$ is contained in a ball of radius $M$. Let $m_t = \mathbb{E}[\mu_{data}]$ and $C_t = Cov[\mu_{data}] + tI_d$ be the optimal mean and covariance operator from Lemma 2. Denote $n_t = \mathcal{N}(m_t, V_t)$. Then $KL(p_t|n_t) \to 0$ as $t \to \infty$.*

*Proof.* For $t > 0$, the KL-divergence is given by

$$\begin{aligned} KL(p_t|n_t) &= -H(p_t) + \frac{1}{2}\log\left(\det(2\pi V_t)\right) + \frac{1}{2}\mathrm{tr}((C+tI_d)V_t^{-1}) \\ &= -H(p_t) + \frac{1}{2}\log\left(\det(2\pi V_t)\right) + \frac{d}{2} \\ &= -H(p_t) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log\left(\prod_{i=1}^{d}(c_i+d)\right) + \frac{d}{2}. \end{aligned} \tag{19}$$

We used that

$$\det(C + tI_d) = \prod_{i=1}^{d}(c_i + t), \tag{20}$$

where $c_i$ are the eigenvalues of $\mathrm{Cov}[\mu_{\text{data}}]$.

Now,

$$-H(p_t) = \mathbb{E}_{X_t}[\log p_t(X_t)] = \mathbb{E}_{X_t}\left[\log\left(\mathbb{E}_{X_0}\left[(2\pi t)^{-d/2}\exp\left(-\frac{1}{2t}\|X_t - X_0\|^2\right)\right]\right)\right].$$

We bound

$$\exp\left(-\frac{1}{2t}\|X_t - X_0\|^2\right) \leq \exp\left(-\frac{1}{2t}(\|X_t\|^2 - 2\|X_t\|M)\right).$$

We write $X_t$ as $X_t = X_0 + \sqrt{t}Z$ where $Z \sim \mathcal{N}(0, I_d)$. Then,

$$\begin{aligned}
&\log\left(\mathbb{E}_{X_0}\left[(2\pi t)^{-d/2}\exp\left(-\frac{1}{2t}\|X_t - X_0\|^2\right)\right]\right) \\
&\leq \log((2\pi t)^{-d/2}) - \frac{1}{2t}(\|X_t\|^2 - 2\|X_t\|M) \\
&\leq -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log(t^d) - \frac{1}{2t}\|X_t\|^2 + \frac{1}{t}(\|M\| + \sqrt{t}\|Z\|)M.
\end{aligned} \tag{21}$$

We use that

$$\mathbb{E}[\|X_t\|^2] = \mathbb{E}[\|X_0\|^2] + t\mathbb{E}[\|Z\|^2] = \mathbb{E}[\|X_0\|^2] + t\mathbb{E}[\|Z\|^2] = m_0^2 + v_0 + td,$$

where $v_0 = \mathbb{E}[\|X_0 - \mathbb{E}[\mu_{\text{data}}]\|^2] = \mathrm{tr}(C_0)$. Since $Z$ is centred and independent of $X_t$, the cross-term $\mathbb{E}[\langle X_t, Z\rangle] = 0$ vanishes. We now take the expectation over the right hand side of (21) to obtain

$$\mathbb{E}[-\frac{1}{2t}(\|X_t\|^2 - \|X_t\|M)] \leq \frac{1}{2t}\left(-td - m_0 - v_0 + 2\sqrt{t}\mathbb{E}[\|Z\|]M + 2M^2\right).$$

Putting it all together we have that

$$KL(p_t|n_t) \leq \frac{1}{2}\log\left(\frac{\prod_{i=1}^{d}(c_i + t)}{t^d}\right) + \frac{1}{2t}\left(-m_0 - v_0 + 2\sqrt{t}\mathbb{E}[\|Z\|]M + 2M^2\right) \to 0,$$

as $t \to \infty$. □

We are now ready to prove Lemma 3.

*Proof.* We again use (19). We now take the derivative of the KL-divergence using De Bruijn's identity for multivariate random variables (see [Rioul, 2010], [Costa and Cover, 1984]):

$$\frac{\mathrm{d}}{\mathrm{d}t}H(p_t) = \frac{1}{2}\mathbb{E}_{p_t}[\|\nabla \log p_t(X)\|^2] = \frac{1}{2t^2}\mathbb{E}[\|X_t - \mathbb{E}[X_0|X_t]\|^2], \tag{22}$$

where we applied Lemma 6. Since $\mathbb{E}[X_0|X_t]$ is the $L^2$-orthogonal projection of $X_0$ to the $\sigma(X_t)$-measurable random variables and $X_t$ is $\sigma(X_t)$-measurable, we have that

$$\frac{1}{2t^2}\mathbb{E}[\|X_t - \mathbb{E}[X_0|X_t]\|^2] = \frac{1}{2t^2}\left(\mathbb{E}[\|X_t - X_0\|^2] - \mathbb{E}[\|X_0 - \mathbb{E}[X_0|X_t]\|^2]\right) = \frac{d}{2t} - \frac{A_t}{2t^2}.$$

For $A_t = \mathbb{E}[\|X_0 - \mathbb{E}[X_0|X_t]\|^2]$ it holds that

$$0 \leq A_t \leq \mathbb{E}[\|X_0\|^2] = \mathrm{tr}(\mathrm{Cov}(\mu_{\text{data}})),$$

since the conditional expectation is a contraction in $L^2$. We use (20) again to arrive at

$$\frac{\mathrm{d}}{\mathrm{d}t}KL(p_t|n_t) = -\frac{1}{2}\left(\frac{d}{t} - \frac{A_t}{t^2}\right) + \frac{1}{2}\sum_{i=1}^{d}\frac{1}{t + c_i} \geq -\frac{1}{2}\frac{d}{t} + \frac{1}{2}\sum_{i=1}^{d}\frac{1}{t + c_i} =: h'(t), \tag{23}$$

where $c_i$ are the eigenvalues of $\text{Cov}[\mu_{\text{data}}] + tI_d$. Integrating the right hand side we obtain

$$h(t) = \frac{1}{2} \log \left( \frac{\prod_{i=1}^{d} (c_i + t)}{t^d} \right)$$

as a possible antiderivative. Define $g(t) = h(t) - KL(p_t|n_t)$. Let us now assume that the support of $\mu_{\text{data}}$ is contained in a ball of radius $M$. Then, by Proposition 1, $\lim_{t\to\infty} KL(p_t|n_t) = \lim_{t\to\infty} h(t) = 0$ and therefore $\lim g(t) \to 0$. Equation (23) implies that $g'(t) \leq 0$ for all $t > 0$. Assume that there is a $s$ and an $\epsilon > 0$ such that $g(s) \leq -\epsilon < 0$. Since $g'(t) \leq 0$ for all $t > 0$ we can then conclude that $g(t) \leq -\epsilon < 0$ for all $t \geq s$. In particular, $g(t) \not\to 0$ as $t \to \infty$, which is a contradiction. Therefore the statement of the Lemma holds for $\mu_{\text{data}}$.

Now let $\mu_{\text{data}}$ be any measure. Let $X$ be a random variable that is distributed according to $\mu_{\text{data}}$ and $Z$ be a standard normal random variable, $Z \sim \mathcal{N}(0, I_d)$. Then $X + \sqrt{t}Z$ has distribution $p_t$. We define $X^M := X1_{\{\|X\| \leq M\}}$ and $X_t^M = X^M + \sqrt{t}Z$. Denote the distribution of $X_t^M$ by $p_t^M$. The distribution of $X^M$ is supported on a ball of radius $M$. Furthermore, $X^M + \sqrt{t}Z \to X + \sqrt{t}Z$ almost surely and therefore $p_t^M \to p_t$ weakly for any $t$. By lower semi-continuity of the KL-divergence,

$$KL(p_t|n_t) \leq \liminf_{M\to\infty} KL(p_t^M|n_t) \leq \liminf_{M\to\infty} \frac{1}{2} \log \left( \frac{\prod_{i=1}^{d} (c_i^M + t)}{t^d} \right), \tag{24}$$

where $c_i^M$ are the eigenvalues of $\text{Cov}[\mu_{\text{data}}^M] + tI_d$. Since $|X^M| \leq |X|$ and $X$ has first and second moments one can apply dominated convergence to see that the covariance matrices converge, $\text{Cov}(\mu_{\text{data}}^M) \to \text{Cov}(\mu_{\text{data}})$. Especially

$$\lim_{M\to\infty} \prod_{i=1}^{d} (c_i^M + t) = \lim_{M\to\infty} \det(\text{Cov}(\mu_{\text{data}}^M)) \to \det(\text{Cov}(\mu_{\text{data}})) = \prod_{i=1}^{d} (c_i + t).$$

Combining this with (24) concludes the proof. □