A Score-Based Deterministic Diffusion Algorithm with Smooth Scores for General Distributions

Karthik Elamvazhuthi[1*], Xuechen Zhang[1*], Matthew Jacobs[2], Samet Oymak[3], Fabio Pasqualetti[1]

[1]University of California, Riverside
[2]University of California, Santa Barbara
[3]University of Michigan, Ann Arbor
karthike@ucr.edu, xzhan394@ucr.edu, majaco@ucsb.edu,oymak@umich.edu,fabiopas@engr.ucr.edu

Abstract

Score matching based diffusion has shown to achieve the state of art results in generation modeling. In the original score matching based diffusion algorithm, the forward equation is a differential equation for which the probability density equation evolves according to a linear partial differential equation, the Fokker-Planck equation. A drawback of this approach is that one needs the data distribution to have a Lipschitz logarithmic gradient. This excludes a large class of data distributions that have compact support. We present a deterministic diffusion process for which the vector fields are always Lipschitz, and hence the score does not explode for probability measures with compact support. This deterministic diffusion process can be seen as a regularization of the porous media equation, which enables one to guarantee long-term convergence of the forward process to the noise distribution. Though the porous media equation is itself not always guaranteed to have a Lipschitz vector field, it can be used to understand the closeness of the output of the algorithm to the data distribution as a function of the time horizon and score matching error. This analysis enables us to show that the algorithm has better dependence on the score matching error than approaches based on stochastic diffusions. Using numerical experiments we verify our theoretical results on example one and two dimensional data distributions which are compactly supported. Additionally, we validate the approach on modified versions of the MNIST and CIFAR-10 data sets for which the distribution is concentrated on a compact set. In each of the experiments, the approach using deterministic diffusion performs better that the diffusion algorithm with a stochastic forward process, when considering the FID scores of the generated samples.

1 Introduction

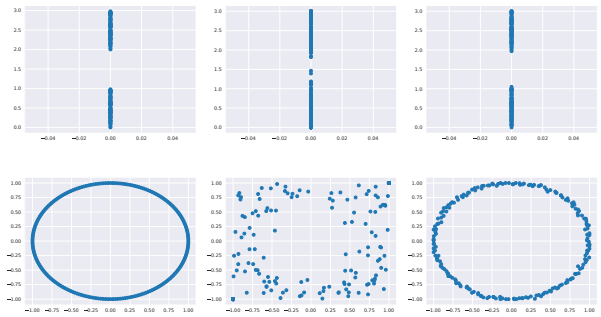
In recent years, score matching based diffusion models have become the state of art in generative modeling (Ho, Jain, and Abbeel 2020; Song et al. 2020; Dhariwal and Nichol 2021). They have found several applications such as image synthesis (Dhariwal and Nichol 2021), protein modeling (Anand and Achim 2022) and inpainting (Lugmayr et al. 2022).

Due to their success, a number of works have focused on the theoretical understanding of the class of data distributions that can be sampled from using score matching based techniques (De Bortoli 2022; Chen et al. 2022; Chen, Lee, and Lu 2023; Lee, Lu, and Tan 2022; Benton, Deligiannidis, and Doucet 2023) study the rate of convergence of solutions of the discrete implementations of the continuous forward and reverse process to the data distribution. A common thread in these works is that if the data distribution has bounded logarithmic gradient, then the distributions can be sampled with error that depends polynomial on some algorithm parameters. To sample from distributions that have a bounded logarithmic gradient implies that the data distribution is positive everywhere on the sample space. For example, if one wants to sample from the set of cats, there is still a small probability that one will sample from the set of chairs. To address this issue, (De Bortoli 2022) study the rate of convergence when the data distribution satisfies the manifold hypothesis that data is concentrated on a lower dimensional manifold. They derived error estimates that scale exponentially with the algorithm parameters using the fact that one can approximate the measure value distribution using a distribution of full support using the regularizing properties of the Fokker Planck equation. Similar ideas have been used to derive early stopping criteria to sample from data distributions that are not fully supported on the sample space (Chen, Lee, and Lu 2023). However, the actual distributions that the approaches sample are still positive everywhere and the exploding score issue cannot be avoided.

Another drawback of stochastic forward processes is that the particle trajectories are not differentiable and the approximation procedure for the reverse phase does not achieve the best error estimates. One way to address this issue is to ensure that the forward phase trajectories are obtained from a deterministic process. For this reason, probabilistic ordinary differential equation (ODE) flows have been analyzed in literature to derive better error estimates (Chen et al. 2023). However, the implementation of the forward algorithm in (Chen et al. 2023) is not strictly deterministic. Similarly, (Liu, Gong, and Liu 2022) constructs deterministic vector fields that transport one measure to another. However, the vector fields can be highly irregular for distributions that are not positive everywhere.

To sample from data general distributions we introduce an alternative deterministic forward process to the probabilistic ODE model. Similar to the probabilistic flow ODE, the vector-field is a function of the particle distribution. How-

*These authors contributed equally.



(a) Target (b) Stochastic (c) Deterministic

Figure 1: Results for distribution with compact support. The initial row illustrates the outcomes for the 1D data distribution, while the subsequent row displays the results for the 2D distribution. For 1D distribution, we set $\Omega_{\text{min}} = 1$ and $\Omega_{\text{max}} = 3$. For 2D distribution, we set $\Omega_{\text{min}} = -1$ and $\Omega_{\text{max}} = 1$.

ever, unlike for the probabilistic ODE model, the evolution of the probability density is given by a nonlinear partial differential equation (PDE). An advantage of using this process is that the vector fields remain bounded and regular for any choice of data distribution. To understand the long-term behavior and approximation properties of this alternative score matching algorithm we study a limit diffusion process as a certain parameter in the ODE goes to zero. In this case, evolution of the probability density of the diffusion process evolves according to a well-known PDE, known as the porous media equation. Using this deterministic approach we also experimentally validate how our algorithm shows superior performance for distributions that are not positive everywhere on the sample space. See Figure 1. Importantly, our empirical findings are in perfect agreement with theory which identifies that deterministic diffusion can be superior when the stochastic diffusion's gradient becomes unstable. Figure 2 follows the setting of Figure 1 and shows that, as time grows and density function evolves to attain near-zero values, stochastic gradient blows up whereas smoothed deterministic gradient remains stable.

The paper is organized as follows. In section 2, we provide background on diffusion based generative modeling. In section 3, we present the deterministic diffusion algorithm presented and studied in this paper. In section 4, we present experiments where we validate the effectiveness of our algorithm over the classical stochastic approach. In section 5 we present some mathematical preliminaries and definitions which are used in the later part of the paper. In section 6, we present analysis on the long-term behavior of the forward process and derive error estimates on the output of the algorithm as a function of the time horizon and the score matching error.

2 Background

Before we discuss the deterministic diffusion based score matching diffusion algorithm presented in the paper, we re-

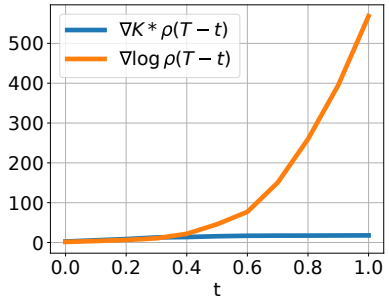


Figure 2: The gradient norm associated with the deterministic diffusion is represented by the blue curve, while the gradient norm corresponding to the stochastic diffusion is depicted in orange. These outcomes were obtained using a 2D distribution with compact support, aligning with the results presented in the second row of Fig. 1.

view the classical score matching based generation as presented in (Song et al. 2020). Let ρ_d be the data distribution from which we desire to sample from. We define the forward process.

$$dX = -\nabla V(X)dt + \sqrt{2}dW + d\psi(t) \qquad (1)$$
$$X_0 \sim \rho_d$$

where $W(t)$ is the standard Brownian motion and ψ is a stochastic process that ensures that the process remains confined to some domain Ω. The probability distribution ρ_d represents to distribution of data, and the goal of generative modeling is to sample from the distribution ρ_d. The potential $V : \mathbb{R}^d \to \mathbb{R}$ is chosen such that the probability density of the random variable $X(t)$ converges to a distribution ρ_n, from which one can easily sample and referred to as the noise distribution. This is guaranteed by analyzing the behavior of the Fokker Planck equation which governs the evolution of $\rho(t)$ given by

$$\frac{\partial \rho}{\partial t} = \Delta \rho + \nabla \cdot (\nabla V(x)\rho) \qquad (2)$$
$$\rho(0) = \rho_d.$$

When Ω is a bounded set, this equation is additionally supplemented by a boundary condition, known as the *zero flux* boundary condition

$$\vec{n}(x) \cdot (\nabla \rho(t,x) + \nabla V(x)) = 0 \text{ on } \partial\Omega \qquad (3)$$

where $\vec{n}(x)$ is the unit vector normal to the boundary of the domain $\partial\Omega$. This boundary condition ensures that preserves that ensures $\int_{\mathbb{R}^d} \rho(t,x)dx = 1$ for all $t \geq 0$. An advantage of considering the situation of bounded domain is that one can choose $V \equiv 0$ and the noise distribution can be taken to be the uniform distribution on Ω. Alternatively, as in (Song et al. 2020), for the choice $V(x) = -\nabla \log \rho_n$ we can show that $\lim_{t\to\infty} \rho(t) = \rho_n$. We can rewrite the above equation as

$$\frac{\partial \rho}{\partial t} = \nabla \cdot ([\nabla \log \rho + \nabla V(x)]\rho) \qquad (4)$$

In order to sample from ρ_d, fixing $T > 0$, one can sample from ρ_n and run the reverse process, also referred to as the *probabilistic flow ODE*,

$$dX = \nabla \log \rho(T-t)dt - \nabla V(x)dt$$
$$X_0 \sim \rho_n \tag{5}$$

However, in practice one does not have complete information about the *score*: $\nabla \log \rho(T-t)$. Therefore, a neural network $s(t, x, \theta)$ is used to approximate this quantity by solving the optimization problem,

$$\min_{\theta} \quad \int_0^T \mathbb{E}_{\rho(T-t)} |s(t, \cdot, \theta) - \nabla \log \rho(T-t)|^2 dt \tag{6}$$

This objective ensures that the solution $\rho_\theta(t)$ of the equation

$$\frac{\partial \rho_\theta}{\partial t} = \nabla \cdot ([s(t, x, \theta) - \nabla V(x)]\rho_\theta) \tag{7}$$
$$\rho_\theta(0) = \rho_n$$

is close to $\rho(T-t)$ so that we can sample from ρ_d by running the reverse ODE,

$$dX = s(t, X, \theta) - \nabla V(x)dt$$
$$X_0 \sim \rho_n \tag{8}$$

3 Deterministic Diffusion Algorithm

One of the drawbacks of this sampling approach from ρ_d is that one requires $\nabla \log \rho(t)$, and hence $\nabla \log \rho_d$ to be bounded, which excludes a large class of probability distributions.

For this reason, we consider an alternative *deterministic* forward diffusion process inspired by the blob method for diffusions, as presented in (Carrillo, Craig, and Patacchini 2019; Craig et al. 2023). Let $K : \mathbb{R}^d \to \mathbb{R}$ be a non-negative *mollifier* function such that $\int_{\mathbb{R}^d} K(x)dx = 1$ and $\lim_{\epsilon \to 0} K_\epsilon(x) := \frac{1}{\epsilon^d} K(\frac{x}{\epsilon}) = \delta_0$, where δ_0 is the Dirac measure at 0.

$$dX(t) = -\nabla K_\epsilon * \rho(t)dt - \nabla V(X)dt + d\psi(t) \tag{9}$$
$$X_0 \sim \rho_d \tag{10}$$

where $\rho(t)$ is the distribution of the variable $X(t)$ and $*$ denotes the convolution operation. An example, of the interaction functional that we take is the Gaussian kernel,

$$K(z) = \frac{1}{(4\pi)^{d/2}} e^{-\frac{|z|^2}{4}} \tag{11}$$

Since the evolution of the process depends on the distribution of the variable itself, this is known as a *Mckean-Vlasov process*. In practice, $\rho(t)$ is approximated by a finite N number of particles $X_i(t)$ for $i = 1, .., N$, and the system (9) is approximated by the system of ODEs

$$dX_i(t) = -\frac{1}{N} \sum_{j=1}^N \nabla K_\epsilon (X_i - X_j)\, dt$$
$$-\nabla V(X_i)dt + d\psi(t)$$
$$X_i(0) \sim \rho_d \tag{12}$$

The term $\frac{1}{N} \sum_{j=1}^N \nabla K_\epsilon (X_i - X_j)$ has a similar effect on the evolution $X_i(t)$ as the noise term $dW(t)$ in (1). However, in the deterministic case particles are diffusing by mutually repelling each other rather than due to the presence of noise. An advantage of using this forward process is that the particle trajectories are deterministic and hence regular as a function of time. This is unlike the stochastic case, where particle trajectories are nowhere differentiable, and hence harder to approximate in the reverse direction. Moreover, for fixed $\epsilon > 0$ the vector-field is always Lipschitz regular. The evolution of the density of the process is then given by

$$\frac{\partial \rho}{\partial t} = \nabla \cdot ([\nabla K_\epsilon * \rho + V(x)]\rho) \tag{13}$$
$$\rho(0) = \rho_d$$

In this case, if $V(x) = \nabla K * \rho_n$, then noise distribution ρ_n becomes an equilibrium point of the above nonlinear partial differential equation. Alternatively, when V is equal to 0 everywhere, the particles will spread close to uniformly on Ω. The long-term behavior for (9) fixed $\epsilon > 0$ is not known, unlike for (1), for which we know that the distribution exponentially converges to equilibrium. However, one can understand the long-term behavior this deterministic diffusion process by considering the limit $\epsilon \to 0$ since $\nabla K_\epsilon * \rho = \nabla \rho$ which is equal to 0 for the uniform distribution. See section 6.

Similar to the stochastic case we define the optimization problem to be solved for the reverse process

$$\min_{\theta} \quad \int_0^T \mathbb{E}_{\rho(T-t)} |s(t, \cdot, \theta) + \nabla K * \rho(T-t)|^2 dt$$

As in the stochastic case, one can sample from the distribution ρ_d by running reverse process,

$$dX = s(t, X, \theta)dt + \nabla V(x)dt + d\psi(t) \tag{14}$$
$$X_0 \sim \rho_n$$

The algorithm for the deterministic diffusion based sampling in presented in Algorithm 1, where we take $V \equiv 0$, so that the noise distribution is the uniform distribution on Ω.

4 Experiments

In this section, we illustrate the advantages inherent in employing deterministic diffusion as opposed to the stochastic forward process. Through the presentation of results, we aim to corroborate our theoretical conclusions. To this end, we conduct the following experiments:

- We compare the performance of the stochastic and deterministic forward processes across data distributions for two a that are have take positive values only in a strict subset of the sample space.
- We show the gradients of the deterministic and stochastic diffusion processes during the reverse phase.
- We compare the performance of the deterministic and stochastic diffusion techniques manifest when applied to real-world datasets binary MNIST (LeCun et al. 1998) and CIFAR-10 (Krizhevsky, Nair, and Hinton 2014).

Initially, we generate 1D and 2D data distributions which have a compact support, which results in taking a value equal to zero for a large portion of the sample space. To create the 1D distributions, we uniformly distribute 200 data points

Algorithm 1: Deterministic

Parameters: # of local epochs M; # of samples N; the range of the distribution $[\Omega_{\min}, \Omega_{\max}]$

Generating Samples from μ_{data}:

$X_0 \rightarrow$ sample from given samples $\{x^i\}_{i=1}^N$

Training

for epoch 1 to M **do**
  $t \rightarrow$ sample from $\mathcal{U}(0, T)$
  for $k = 1$ to t **do**
    for i in range(N) **do**
      $z_i \rightarrow$ sample from $\mathcal{N}(0,\, 1)$
      $X_i^{k+1} = X_i^k + \Delta t \frac{1}{N} \sum_{j=1}^N \nabla K_\epsilon \left(X_i - X_j\right)$
      project X_i^{k+1} to Ω
    end for
  end for
  $l(\theta, t) = \frac{1}{N} \sum_{i=1}^N |s(X_i, t, \theta) + \frac{1}{N^2} \sum_{j=1}^N \nabla K_\epsilon (X_i t - X_j t)|^2$
  $\theta = \text{optimizer\_step}(l(\theta))$
end for

Sampling

$Y_0 \rightarrow$ sample from uniform distribution
for k in range(T) **do**
  $Y^{k+1} = Y^k + \Delta t s_\theta(Y^k, k)$,
  project Y^{k+1} to Ω
end for

Function: project X_i^{k+1} to Ω

for $ii = 0$ to 1 **do**
  if $X_i^{k+1}[ii] > \Omega_{\max}$ **then**
    $X_i^{k+1}[ii] = \Omega_{\max}$
  end if
  if $X_i^{k+1}[ii] < \Omega_{\min}$ **then**
    $X_i^{k+1}[ii] = \Omega_{\min}$
  end if
end for

within the intervals [0,1] and [2,3]. Outside of these intervals, no data points are present. To generate a 2D data distribution, we create a circle of radius one centered at (0,0). In total, we generate 200 points uniformly distributed along the circumference of this circle.

In another example, we also generate 1D and 2D distributions that are positive in the entire sample space. To obtain the 1D distribution, we uniformly distribute 200 data points within the intervals [1,2] and [3,4], and additionally introduce 200 extra points uniformly distributed within the range [0,5]. Similarly, for the 2D distribution, we augmented the existing circular distribution with an additional 200 points uniformly distributed within the specified range[-2,2].

To test the effectiveness of the approach for different choices of the functions K and V, we ran the experiments with compactly supported data distribution, for the two dimensional synthetic data for two different choices of kernel functions, $K_1 = \exp(\frac{-1}{1-|x/\epsilon|^2})\mathbf{1}_{|x|<1}, K_2 = (1 - |x/\epsilon|^2)^2\mathbf{1}_{|x|<1})$ and Gaussian noise distribution V.

Because the distributions are relatively straightforward and comprise only a limited number of samples, we employ a basic 4-layer fully connected model with ReLU as the activation function for our diffusion model.

In terms of implementation specifics, we utilize a batch size of 16 and train for 500 epochs. The optimization uses the Adam optimizer with a learning rate of 1e-3. Additionally, we set the value of Δ_t to 0.01, implying a total of 100 steps in the process.

We provide a visualization of the generated distributions when the data distributions have a compact support in Fig. 1. Remarkably, the deterministic diffusion technique exhibits notable enhancement in performance. In contrast, the original stochastic diffusion approach encounters limitations in accommodating general distributions since the vector fields are not bounded in this case. Our algorithm, incorporating deterministic forward processes, adeptly handles such scenarios. Concerning distributions featuring full support on Ω, both diffusion methods demonstrate comparable performance. The results are shown in Fig. 5. The results of the experiments for different choices of kernel and noise distribution are shown in Figure 3 and 4. Our empirical findings reveal a consistent trend of improvement in the deterministic algorithm's performance across these various kernel and noise distribution choices. This consistency underscores the robustness of our approach. The first kernel function we explored exhibited similar performance to the Gaussian kernel. This indicates that a judicious choice of the kernel might result in a better algorithm.

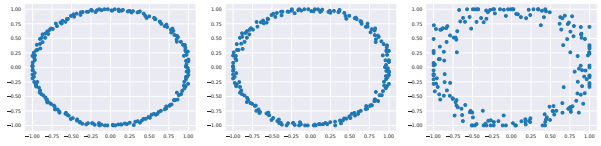
In addition to the synthesized toy examples, we extend our evaluation to real-world datasets, specifically binary MNIST images, for which the data distribution does not exhibit positivity throughout the sample space. The outcomes are illustrated in Fig. 6. Notably, the deterministic diffusion method continues to exhibit discernible enhancements in performance.

We calculated the Wasserstein distance between the generated data distribution and the target distribution to facilitate a more straightforward assessment of the diffusion model's performance. The results are presented in Table 1.

We conducted simulations of the reverse phase on the 1D distribution featuring compact support, employing both deterministic and stochastic diffusion methods. We performed a numerical evaluation of the gradient for different values of t. Notably, as $t \rightarrow 1$, the gradient remains stable for deterministic diffusion, without any indications of exploding behavior. Conversely, for the stochastic approach, the gradient does exhibit an exploding tendency.

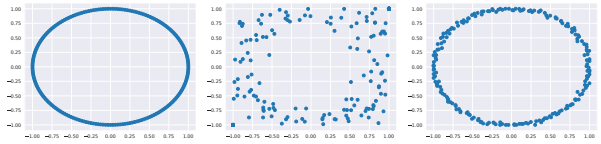
We additionally conducted experiments on the CIFAR-10 dataset. To highlight the effect of compact support of the data distribution, we extended the original 32×32 pixel images to 40×40 pixels by padding with black pixels. Subsequently, we trained our model on both stochastic and deterministic approaches.

We compared the FID scores specifically on the original 32×32 pixel portion of the generated images, and the results are presented in Table 1. Our findings indicate that the deterministic approach performs significantly better than the stochastic approach in terms of FID scores. This empirical evidence demonstrates the effectiveness of our method,



(a) K_{Gaussian} (b) K_1 (c) K_2

Figure 3: Results on 2D data set for different choices of the Kernel.



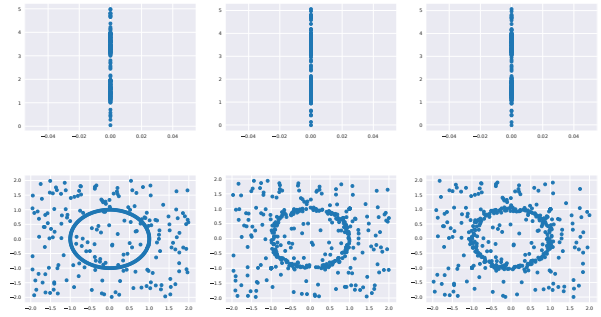
(a) Target (b) Stochastic (c) Deterministic

Figure 4: Results on 2D data set when the noise distribution is Gaussian.

particularly when applied to more substantial and complex datasets like CIFAR10.

5 Mathematical Preliminaries and Definitions

In this section, we define some notation that will be used in section 6 where we present where we will present some analysis about the algorithm presented in 3. We refer the readers to (Ambrosio, Gigli, and Savaré 2005) for more details. Let $\Omega \subset \mathbb{R}^d$ be a convex set. Without loss of generality we assume that Ω has Lebesgue measure 1. Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the set of Borel probability measures on $\mathbb{R}^d$ with finite second moment: $\int_\Omega |x|^2 d\mu(x) < \infty$. For a given Borel map



(a) Target (b) Stochastic (c) Deterministic

Figure 5: Results for distribution without compact support. The initial row illustrates the outcomes for the 1D data distribution, while the subsequent row displays the results for the 2D distribution. For 1D distribution, we set $\Omega_{\text{min}} = 0$ and $\Omega_{\text{max}} = 5$. For 2D distribution, we set $\Omega_{\text{min}} = -2$ and $\Omega_{\text{max}} = 2$.

Positivity	Distribution	Deterministic	Stochastic
Yes	1D	2.315	2.290
	2D	3.873	3.964
No	1D	1.953	3.143
	2D	2.057	4.628
	Bin MNIST	15.61	23.27
	CIFAR-10	27.93	42.05

Table 1: Wasserstein distances. Average of 5 runs

$T : \mathbb{R}^d \to \mathbb{R}^d$ we will denote by $T_\#$ the corresponding push-forward map, which maps any measure μ to a measure $T_\#\mu$, where $T_\#\mu$ is the measure defined by

$$(T_\#\mu)(B) = \mu(T^{-1}(B)), \tag{15}$$

for all Borel measurable sets $B \subseteq \mathbb{R}^d$. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we denote the set of transport plans from μ to ν by

$$\Gamma(\mu, \nu) := \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) | \pi^1_\#\gamma = \mu, \pi^2_\#\gamma = \nu\}, \tag{16}$$

where $\pi^i : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ are the projections on to the ith coordinates, respectively. We will define the $2-$Wasserstein distance between two probability measures μ, ν as the following

$$W_2(\mu, \nu) = \min_{\gamma \in \Gamma(\mu,\nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 d\gamma(x, y) \right)^{1/2}. \tag{17}$$

Suppose that K is smooth. We define the functional $\mathcal{E}_\epsilon$ given by

$$\mathcal{E}_\epsilon(\mu) = \frac{1}{2} \int_{\mathbb{R}^d} \int_{\mathbf{R}^d} K_\epsilon(x - y)\, d\mu(y) d\mu(x) \tag{18}$$

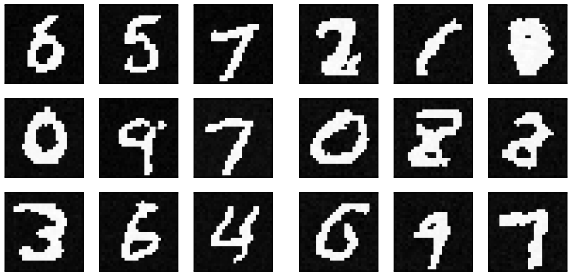
for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$

Additionaly, we define the functional

$$\mathcal{V}(\mu) = \int_{\mathbb{R}^d} V(x)\mu(x)dx \tag{19}$$

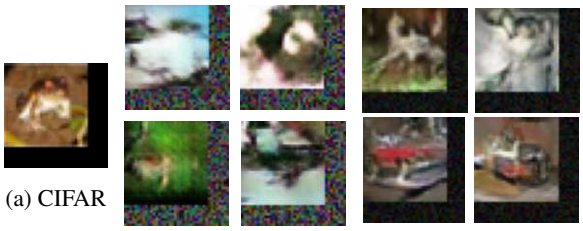
where $V \in C^2(\mathbb{R}^d)$ be a smooth strongly convex function. That is, there exists $\lambda > 0$ such that

$$V(\gamma x + (1-\gamma)y) \leq \gamma V(x) + (1-\gamma)V(y) + \frac{\lambda\gamma(1-\gamma)}{2}\|x - y\|^2$$



(a) deterministic (b) stochastic

Figure 6: Binary MNIST



(a) CIFAR (b) Stochastic (c) Deterministic

Figure 7: Unconditional CIFAR10 generation.

for all $x, y \in \mathbb{R}^d$ and all $\gamma \in (0,1)$, such that $-V$ is positive everywhere and $-\int_\Omega V(x)dx = 1$. Additionally, we will define the functional

$$\mathcal{V}_\Omega(\mu) = \begin{cases} 0 \text{ if supp } \mu \subseteq \Omega \\ \infty \text{ otherwise} \end{cases} \tag{20}$$

It follows from the theory of gradient flows on Wasserstein spaces (Ambrosio, Gigli, and Savaré 2005), that the PDE (13), can be expressed as a gradient flow of the functional $\mathcal{F} := \mathcal{E}_\epsilon + \mathcal{V} + \mathcal{V}_\Omega$. Similarly, the PDE can be expressed as gradient flow of the functional $\mathcal{E}$.

$$\frac{\partial \rho}{\partial t} = -\nabla_{W2}\mathcal{F} \tag{21}$$

$$= \nabla \cdot (\rho \nabla \frac{\partial \mathcal{F}}{\partial \rho}) \tag{22}$$

where formally ∇_{W2} is the Wasserstein gradient of $\mathcal{F}$.

In order to understand the long-term behavior and the performance of the algorithm, we introduce another forward process by taking the limit $\epsilon \to 0$ to get

$$dX(t) = -\nabla\rho(t)dt - \nabla V(X)dt + d\psi(t) \tag{23}$$

$$X_0 \sim \rho_d \tag{24}$$

In this case the evolution of the probability density function is given by,

$$\frac{\partial \rho}{\partial t} = \nabla \cdot ([\nabla \rho + V(x)]\rho) \tag{25}$$

$$\rho(0) = \rho_d \tag{26}$$

This PDE can also be seen as the gradient flow of the functional $\mathcal{F} := \mathcal{E} + \mathcal{V} + \mathcal{V}_\Omega$. where functional $\mathcal{E} : \mathcal{P}_2(\Omega) \to [0, \infty)$ is given by,

$$\mathcal{E}(\mu) = \begin{cases} \frac{1}{2}\int_\Omega |\mu(x)|^2 dx \text{ if } \mu << \mathcal{L} \\ \infty \text{ otherwise} \end{cases} \tag{27}$$

where $\mathcal{L}$ denotes the Lebesgue measure on $\mathbb{R}^d$ and $\mu << \mathcal{L}$ denotes that μ is absolutely continuous with respect to the Lebesgue measure. This introduces the score matching optimization problem for the limit $\epsilon \to 0$ given by

$$\min_\theta \quad \int_0^T \mathbb{E}_{\rho(T-t)} |s(t,\cdot,\theta) + \nabla\rho(T-t)|^2 dt$$

We will quantify in the next section the closeness of $\rho_\theta(T)$ to ρ_d for the equation

$$\frac{\partial \rho_\theta}{\partial t} = \nabla \cdot ([s(t,x,\theta) - \nabla V(x)]\rho_\theta) \tag{28}$$

$$\rho_\theta(0) = \rho_n$$

6 Analysis

In this section, we present some analysis on the long-term behavior of (13) by studying the behavior of (25). We first verify that for the limit forward process $\epsilon = 0$ the distribution of the process converges to the noise distribution as $t \to \infty$.

Lemma 6.1. *Let ρ be the gradient flow of the functional $\mathcal{F}$. Then we have that*

$$\lim_{t\to\infty} W_2(\rho(t), \rho_n) \leq e^{-\lambda t} W_2(\rho_T, \rho_n) \tag{29}$$

where $\rho_n = -V$.

Suppose V_n is only convex and not necessarily strongly convex, then

$$\lim_{t\to\infty} W_2(\rho(t), \rho_n)) = 0 \tag{30}$$

where $\rho_n = -V$ if V is non-zero and equal to $c\mathbf{1}_\Omega$ otherwise, where $\mathbf{1}_\Omega$ is the characteristic function of Ω and $c > 0$ is a constant such that $c\mathbf{1}_\Omega$ is the uniform distribution on Ω.

Proof. The functional $\mathcal{F}$ is $\lambda-$*convex on* $\mathcal{P}_2(\mathbb{R}^d)$ *along generalized geodesics* (See (Ambrosio, Gigli, and Savaré 2005)) due to the strong convexity of the function V. Hence, the exponential convergence result is well known due to (Ambrosio, Gigli, and Savaré 2005)[Theorem 11.2.1].

In the case when V is not strongly convex, the functional $\mathcal{E}$ is convex on $\mathcal{P}_2(\Omega)$, but not λ-convex for $\lambda > 0$. We note that the functional $\mathcal{E}$ has the same minimizers as

$$\Gamma(\mu) = \mathcal{F}(\mu) + \int_{\mathbb{R}^d} V^2(x)dx \tag{31}$$

The functional $\frac{1}{2}\int_\Omega |\mu(x) + V(x)|^2 dx$ is strictly convex in $L^2(\Omega) \cap \mathcal{P}_2(\Omega)$, the set of square integrable functions on Ω and hence has a unique global minimizer $-V(x)$. Hence, on measures with support in Ω, the functional $\mathcal{F}$ has the same minimizers as the functional $\mu \mapsto \frac{1}{2}\int_\Omega |\mu(x) + V(x)|^2 dx + \mathcal{V}_\Omega$. Therefore, the global minimizer in $\mathcal{P}_2(\Omega)$ is also unique. Then the result follows from (Ambrosio, Gigli, and Savaré 2005)[Corollary 4.0.6]. □

Though, for $V = 0$ we do not have exponential stability in the Wassertstein-2 metric, it is known to be exponentially convergent in the case of a bounded domain in the L^∞ norm for long time.

Lemma 6.2. *(Grillo and Muratori 2013) Let $V \equiv 0$. Let $\rho(t,x)$ be the solution of* (25)*. Then there exists $C, m > 0$*

$$\|\rho(t) - \mathbf{1}_\Omega\|_\infty < Ce^{-mt}$$

for all $t \geq 1$.

Unfortunately, due to the lack convexity of the functional $\mathcal{F}_\epsilon$ on $\mathcal{P}_2(\mathbb{R}^d)$ it is not possible to use the same argument to infer the long term convergence of (13) to the noise distribution. However, we know the following.

Lemma 6.3. *The minimizers of $\mathcal{F}_\epsilon$ converge to the minimizers of $\mathcal{F}$ in $\mathcal{P}_2(\mathbb{R}^d)$ as $\epsilon \to 0$.*

Proof. In (Craig et al. 2023)[Theorem 5.1] it has been shown that the functional $\mathcal{E}_\epsilon + \mathcal{V}$ *gamma converges* to the functional $\mathcal{E} + \mathcal{V}$ as $\epsilon \to 0$. Firstly, this means that, for every sequence $\mu_\epsilon \in \mathcal{P}_2(\mathbb{R}^d)$ converging to $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathcal{E}(\mu) + \mathcal{V}(\mu) \leq \liminf_{\epsilon \to \infty} \mathcal{E}_\epsilon(\mu_\epsilon) + \mathcal{V}_\epsilon(\mu_\epsilon)$$

Hence, we can also conclude that

$$\mathcal{F}(\mu) \leq \liminf_{\epsilon \to \infty} \mathcal{F}_\epsilon(\mu_\epsilon)$$

Secondly, from the gamma convergence result we know that for every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists a sequence $\mu_\epsilon \in \mathcal{P}_2(\mathbb{R}^d)$ such that

$$\mathcal{E}(\mu) + \mathcal{V}(\mu) \geq \limsup_{\epsilon \to \infty} \mathcal{E}_\epsilon(\mu_\epsilon) + \mathcal{V}_\epsilon(\mu_\epsilon).$$

From the definition of $\mathcal{V}_\Omega$, this also implies that

$$\mathcal{F}(\mu) \geq \limsup_{\epsilon \to \infty} \mathcal{F}_\epsilon(\mu_\epsilon).$$

Hence, $\mathcal{F}_\epsilon$ gamma converges to $\mathcal{F}$ as $\epsilon \to \infty$. This along with coercivity of the functionals $\mathcal{F}_\epsilon$, implies that minimizers of $\mathcal{F}_\epsilon$ converge to minimizers of $\mathcal{F}$. □

The previous result, however, does not immediately give as that the solution ρ_ϵ of (13) are converging to solutions of (25) as $\epsilon \to 0$. However, it has been shown in (Craig et al. 2023), that one can find a sequence of regularizations $\mathcal{F}_{\epsilon,k}$ of the functional $\mathcal{F}_k$ for which, the corresponding gradient flows converge to that of $\mathcal{F}$ for a suitable sequence of k, ϵ. This is achieved by regularizing the functional $\mathcal{V}_\Omega$.

Next, we look at the approximation capabilities of the score matching algorithm as a function of the time horizon and the score matching error, under the assumption that the actual score and the approximating one are Lipschitz. This is a common assumption in error analysis of score matching algorithms (Chen, Lee, and Lu 2023). Since the processes are deterministic we are able to get linear dependence of the error in density estimation as a function of the score error. This is unlike the stochastic case, where the dependence is a function of the square root of the score error.

Lemma 6.4. *Suppose V is strongly convex with constant $\lambda > 0$ and $\bar{\Omega}$ is compact. Let $\rho \in C^1([0,T] \times \bar{\Omega})$ be a solution of $\mathcal{E}$. Suppose $\nabla\rho(t)$ and $s(t\cdot, \theta)$ is uniformly Lipschitz for a Lipschitz constant L. Moreover, assume that the support of $\rho(t)$ lies in a compact set K for all $t \geq 0$ and*

$$\int_0^T \mathbb{E}_{\rho(T-t)} |s(t,\cdot,\theta) + \nabla\rho(T-t)|^2 dt < \epsilon^2 \tag{32}$$

Then

$$W_1(\rho_\theta(T), \rho_d) \leq Ce^{(L-\lambda)T} W_2(\rho_d, \rho_n) + \epsilon e^{LT}.$$

Proof. Due to the assumptions of Lipschitnzess, we can conclude from Theorem (Benton, Deligiannidis, and Doucet 2023)[Theorem 1] that the flows satisfy the estimate

$$W_2(\tilde{\rho}(T), \rho_\theta(T)) < \epsilon e^{LT} \tag{33}$$

where $\tilde{\rho}$ is the solution of the equation

$$\frac{\partial \tilde{\rho}}{\partial t} = \nabla \cdot (-\nabla\rho(T-t) - \nabla V(x)\tilde{\rho}) \tag{34}$$

$$\rho(0) = \rho_n \tag{35}$$

Due to the Lipschitz assumption on $\nabla\rho(T-t)$ we can estimate the distance between the the final conditions as a function of the distance between the initial conditions. From (Bonnet and Frankowska 2021)[Proposition 2] we can conclude that

$$W_2(\tilde{\rho}(T), \rho_d) \leq Ce^{LT} W_2(\rho_n, \rho(T)) \tag{36}$$

Using the triangle inequality for the Wasserstein-2 distance, we can conclude that

$$W_2(\rho_\theta(T), \rho_d) < W_2(\tilde{\rho}(T), \rho_\theta(T)) + W_2(\tilde{\rho}(T), \rho_d)$$

Applying the estimate from (33) and (36) it follows from Lemma 6.1 and (33) that

$$W_2(\rho_\theta(T), \rho_d) < \epsilon e^{LT} + Ce^{LT} W_2(\rho(T), \rho_n)$$

Then applying the result in Lemma 6.1 to bound the second term, the result follows. □

7 Conclusion

We presented a deterministic diffusion algorithm for generative modeling. Experiments show that the presented algorithm performs much better than the original score matching algorithm based on a stochastic forward process. This behavior is due to the bounded score for the distribution for any kind of data distribution. In contrast, the score explodes in value for the stochastic approach for the cases when the distribution has a compact support inside the domain. In addition, we justify the experimental results based on analysis of long-term behavior of the forward process and its approximation properties. Due to the determinism in the trajectories we are able to get linear dependence of the error in density estimation as a function of the score error.

Acknowledgements

This work was supported by the awards AFOSRFA9550-19-1-0235, NSF grants CCF-2046816 and CCF-2212426, Google Research Scholar award, Adobe Data Science Research award, and Army Research Office grant W911NF2110312.

References

Ambrosio, L.; Gigli, N.; and Savaré, G. 2005. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Anand, N.; and Achim, T. 2022. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*.

Benton, J.; Deligiannidis, G.; and Doucet, A. 2023. Error Bounds for Flow Matching Methods. *arXiv preprint arXiv:2305.16860*.

Bonnet, B.; and Frankowska, H. 2021. Differential inclusions in Wasserstein spaces: the Cauchy-Lipschitz framework. *Journal of Differential Equations*, 271: 594–637.

Carrillo, J. A.; Craig, K.; and Patacchini, F. S. 2019. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58: 1–53.

Chen, H.; Lee, H.; and Lu, J. 2023. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, 4735–4763. PMLR.

Chen, S.; Chewi, S.; Lee, H.; Li, Y.; Lu, J.; and Salim, A. 2023. The probability flow ODE is provably fast. *arXiv preprint arXiv:2305.11798*.

Chen, S.; Chewi, S.; Li, J.; Li, Y.; Salim, A.; and Zhang, A. R. 2022. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*.

Craig, K.; Elamvazhuthi, K.; Haberland, M.; and Turanova, O. 2023. A blob method for inhomogeneous diffusion with applications to multi-agent control and sampling. *Mathematics of Computation*.

De Bortoli, V. 2022. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Grillo, G.; and Muratori, M. 2013. Sharp short and long time L_∞ bounds for solutions to porous media equations with homogeneous Neumann boundary conditions. *Journal of Differential Equations*, 254(5): 2261–2288.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Krizhevsky, A.; Nair, V.; and Hinton, G. 2014. The CIFAR-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55(5).

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Lee, H.; Lu, J.; and Tan, Y. 2022. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35: 22870–22882.

Liu, X.; Gong, C.; and Liu, Q. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.