Final report:
A Variational Perspective on Diffusion-Based Generative Models and Score Matching

Imane Elbacha, Nathan Weill
MVA 22 Generative Models

Abstract

The paper "A Variational Perspective on Diffusion-Based Generative Models and Score Matching" [1] proposes a novel variational framework for estimating the marginal likelihood of continuous-time diffusion models, which has significant implications for the understanding of generative modeling techniques. Inspired by Song et al. [2]'s work on discrete-time diffusion-based generative models and score matching methods, this paper addresses the theoretical underpinnings of generative diffusion and provides a unified approach to studying different types of generative models.

In this report, we explore the theoretical underpinnings of this variational framework and its implications for generative modeling techniques. We examine how this framework provides a theoretical foundation for these models and allows for more accurate modeling of high-dimensional image data. Additionally, we discuss how this paper highlights the importance of likelihood estimation in generative models and how it advances our understanding of these techniques. We analyze how this variational perspective can be used to improve our understanding of the underlying mechanisms behind score-based generative models.

Our contribution in this project is centered around the following question: What is the importance of a variational framework? What does it imply theoretically speaking? We also replicated the numerical experiments on other datasets.

1. Introduction

Deep generative modeling techniques have gained popularity in recent years owing to their ability to produce realistic data samples, with results that are increasingly successful. Among these techniques, score-based generative modeling involves learning the gradient of the log-density of perturbed data to define a generative diffusion process. Despite its empirical success, a theoretical foundation for this procedure is yet to be established.

In order to address this gap in knowledge, the research paper [1] proposes a variational framework for estimating the marginal likelihood of continuous-time diffusion models. The proposed framework enables the study of a wide range of models, such as continuous-time normalizing flows and score-based generative models. In fact, through maximizing a lower bound on the marginal likelihood, this framework provides a theoretical basis for these models, enabling more precise modeling of high-dimensional image data.

The importance of this variational framework lies in its ability to provide insight into the relationship between various types of generative models and to advance our understanding of generative modeling techniques. By deriving a general variational framework for likelihood estimation, the paper [1] provides a theoretical underpinning for these models and allows for more accurate modeling of high-dimensional image data. Additionally, this framework provides valuable theoretical insight into how score-based generative models can be studied within a broader class of continuous-time diffusion models.

In this report, we delve into the findings, theoretical foundation, and proposed research framework of the paper. Specifically, we aim to analyze how this framework offers a unified approach to studying diverse types of generative models and how it can enhance our comprehension and implementation of these techniques. Our primary objective is to investigate the significance of the variational framework and its ability to bridge the gap between the empirical success of generative techniques and their foundational theory. With this focus, we aim to highlight the added value of this framework in advancing our understanding of generative models.

2. Paper analysis

2.1. The plug-in reverse SDE

The paper [1] provides an overview of diffusion-based generative models and score matching methods. These techniques have shown promise in modeling high-dimensional image data. The paper builds on previous work by Song et al. [2], who proposed a generative diffusion process that can be reversed by learning the score function, which is the gradient of the log-density of perturbed data.

For the rest of this report, we will adopt the convention of the paper: Y will be the inference or forward process, whereas X will be the generative or backward process.

To derive a theoretical underpinning for this approach, the authors assume that Y_0 follows the data distribution $q(y, 0)$ and Y_s satisfies an Itô SDE given by:

$$\mathrm{d}Y = f(Y, s)\mathrm{d}s + g(Y, s)\mathrm{d}\hat{B}_s$$

where f and g are chosen such that the density $q(y, s)$ will converge to some tractable prior p_0 as $s \to T$. The authors then follow in the footsteps of [2], reminding the reader that it is possible to find a "reverse" SDE whose marginal density evolves according to $q(y, s)$ reversed in time. It writes as:

$$\mathrm{d}X = (gg^\top \nabla \log q(X, T - t) - f)\mathrm{d}t + g\mathrm{d}B_t \quad (1)$$

This provides a foundation for their study and allows them to derive a variational framework for likelihood estimation.

Furthermore, the importance of the score function is emphasized in the approach of the paper. The score function is defined as the gradient of the log-density of perturbed data $\nabla \log q$, which is used to learn the generative model. The proposed method relies on approximating the score via a parameterized score function s_θ by minimizing an integral that involves an explicit score matching (ESM) loss L_{ESM}. The function to minimize is:

$$\int_0^T \mathbb{E}_{Y_s}\left[\frac{1}{2}\left\|\mathbf{s}_\theta(Y_s, s) - \nabla \log q(Y_s, s)\right\|^2_{\Lambda(s)}\right]\mathrm{d}s$$

However, since access to the ground truth score $\nabla \log q$ is not available, alternative losses such as implicit score matching (ISM), sliced score matching (SSM), and denoising score matching (DSM) can be used instead. These alternative losses are all equal up to a constant and can be used interchangeably in practice. By learning the score function, the authors are able to obtain a generative model that can be used for image generation tasks.

They call the new equation, where the score function has been replaced by the learnt parametrized (in practice a neural network) s_θ the plug-in reverse SDE and it writes as follows:

$$\mathrm{d}X = (gg^\top s_\theta - f)\mathrm{d}t + g\mathrm{d}B_t \quad (2)$$

2.2. Continuous-time ELBO

The strength of the paper lies in its derivation in continuous time of the evidence lower bound (ELBO). In fact the authors derive such an ELBO for a general stochastic diffusion equation defined by:

$$\mathrm{d}X = \mu(X, t)\mathrm{d}t + \sigma(X, t)\mathrm{d}B_t$$

where B is a Wiener process and X_t has a density $p(, t)$ with a given initial density p_0. This allows for maximum-likelihood calculations for any diffusion equation.

Then they use three well-known properties of SDEs to compute a lower bound of $\log p(x, t$.

The Fokker-Planck equation gives a partial differential equation satisfied by the density p.

The Feynman-Kac property allows the authors to represent the solution of the EDP p as the expectation of a certain function of stochastic process Y. Y solves

$$\mathrm{d}Y = -\mu(Y, T - s)\mathrm{d}s + \sigma(T - s)\mathrm{d}B'_s$$

It yields

$$p(x, T) = \mathbb{E}\left[p_0(Y_T)\exp(\int_0^T -\nabla\mu(Y_s, T - s)\mathrm{d}s)|Y_0 = x\right]$$

The resulting log likelihood is intractable, so the authors apply a change-of-measure, introducing a variational approximation $\mathbb{Q}$ to $\mathbb{P}$. Girsanov theorem is used to express the Radon-Nikodym derivative $\frac{d\mathbb{Q}}{d\mathbb{P}}$. Then, they obtain the following ELBO after using Jensen's inequality:

$$\mathbb{E}\left[\int_0^T -(\frac{1}{2}\left\|a(\omega, s)\right\|_2^2 + \nabla\mu)\mathrm{d}s + \log p_0(Y_T)\right] := \mathcal{E}^\infty \quad (3)$$

where

$$\mathrm{d}Y = (-\mu + \sigma a)\mathrm{d}s + \sigma\mathrm{d}\hat{B}_s \quad (4)$$

and

$$\mathrm{d}\hat{B}_s = -a(\omega, s)\mathrm{d}s + \mathrm{d}B'_s$$

a must respect the so-called Novikov condition. We call equation (4) the inference SDE.

2.3. Score-based generative modeling revisited

Now that we are equipped with this evidence lower bound, we can apply the formula (3) to the plug-in reverse SDE (2) with $a = g^\top s_\theta$. It thus yields a continuous ELBO for our problem:

$$\mathcal{E}^\infty = \mathbb{E}_{Y_T}\left[\log p_0(Y_T)|Y_0 = x\right] - \int_0^T \mathbb{E}_{Y_s}\left[\frac{1}{2}\left\|\mathbf{s}_\theta)\right\|^2_{gg^\top} + \nabla(gg^\top \mathbf{s}_\theta - f)|Y_0 = x\right]\mathrm{d}s \quad (5)$$

This indicates that matching the score is equivalent to maximizing the evidence lower bound of the plug-in reverse SDE since the ELBO corresponds to the implicit score-mathcing loss up to a constant and with a specific weighting ($\Lambda = gg^\top$).

In fact the authors prove a stronger result by considering a family of plug-in reverse SDEs indexed by a parameter $\lambda \leq 1$:

$$\mathrm{d}X = ((1 - \frac{\lambda}{2})g^2\mathbf{s}_\theta - f)dt + \sqrt{1-\lambda}g\mathrm{d}B_t \qquad (6)$$

and

$$\mathrm{d}Y = (f - \frac{\lambda}{2})g^2\nabla \log q\mathrm{d}s) + \sqrt{1-\lambda}g\mathrm{d}\hat{B}_t \qquad (7)$$

The two limiting cases are for $\lambda = 0$ when we find the original equation and the case $\lambda = 1$ for which the reverse process becomes a deterministic ODE (also called continuous-time flow) since the drift term vanishes.

They then show that the average CT-ELBO of the λ-plug-in reverse SDE is also equivalent ISM loss, up to some constants. Therefore when matching the loss, we maximize in fact the likelihood of the whole family (meaning for every λ) of plug-in reverse SDEs.

Thus if the model has been trained to match the score, we can use the obtained score function $\mathbf{s}_\theta$ to sample from the plug-in reverse SDEs with different values of λ with the Euler-Maruyama scheme. We can also monitor the negative ELBO during training for different values of λ to compare their computation-estimation trade-off.

2.4. Analysis of the implication of these results

These theoretical results do not explain the sample of quality of Denoising Diffusion Probabilistic Models, which could be rather explained by convergence results such as [3]. But they achieve a connection of score-matching and maximum likelihood of the plug-in reverse SDEs family.

We can draw two links with what we saw in class, and with this concurrent work [4] that also derives a lower bound on the likelihood of the plug-in reverse SDE in a different fashion.

We recall from class that for every plug-in reverse SDE, we can write an equivalent ODE formulation (called continuous time flow or continuous normalizing flow), with the same marginal density. The former dynamics is stochastic, the latter is deterministic. This is also reminded in the third paragraph of [4]. The ODE admits a tractable log-density so we could perform maximum likelihood directly on this formula. However, it would be computationally prohibitive since it involves numerically solving an ODE at each iteration of the optimization. Instead, we turn to the log density of the SDE. It is not directly tractable but the authors of [1] have found a variational lower bound that can be readily computed via Monte Carlo sampling (importance sampling is additionally necessary to debias the estimate). It is noisier but much faster than numerical integration. This provides in practice an efficient and scalable training method.

This is complementary to the remark of the paper on the computation trade-off between continuous-time flows and plug-in reverse SDEs neg-likelihood evolution during training. Having a deterministic model is useful for likelihood computation, interpolation and temperature scaling.

Secondly, we know it is possible to derive an evidence lower bound in the discrete framework of the diffusion model. How do both results compare? We can explicitly use a Gaussian variational distribution as seen in the first class. What is interesting is that while implicit score matching (ISM) loss arises from a variational lower bound on the log likelihood as shown in this paper, when we start with a discrete time argument in DDPM to derive a variational lower bound we can show that what arises is the denoising score matching (DSM) loss. This is echoed in [4] where they also find a connection with the DSM loss.

3. Experiments

To comprehensively assess and test the proposed model, we conducted multiple experiments using three academic datasets. In this section, we examine the model's implementation, discuss the datasets utilized, and evaluate the results obtained.

3.1. Model implementation

3.1.1 Variance Preserving Stochastic Differential Equation (SDE)

The VariancePreservingSDE class represents a stochastic differential equation (SDE) based on the variance-preserving model proposed by Song et al. in 2021 [**?**]. This model aims to preserve the variance of the underlying process while providing a flexible framework for sampling from the SDE. The class encapsulates the mathematical formulation of the SDE and provides methods to compute various parameters.

The key elements of the VariancePreservingSDE class are as follows:

- **Parameters**: The SDE is defined by several parameters, including β_{min}, β_{max}, T, and t_ϵ. These parameters control the behavior of the SDE and influence its drift and diffusion functions.

- **Drift and Diffusion Functions**: The drift function $f(t, y)$ calculates the rate of change of the process at a given time t and value y. It is defined as $f(t, y) = -0.5\beta(t)y$, where $\beta(t) = \beta_{\text{min}} + (\beta_{\text{max}} - \beta_{\text{min}})t$.

The diffusion function $g(t, y)$ determines the randomness or noise in the process and is defined as $g(t, y) = \sqrt{\beta(t)}$.

- **Sampling**: The sample(t, y_0, return_noise=False) method generates samples from the SDE given an initial value y_0 at time t. It utilizes the variance-preserving properties to compute the mean and standard deviation of the stationary distribution. By combining these statistics with Gaussian noise, it generates a sample y_t from the SDE. Optionally, it can also return the noise, standard deviation, and diffusion term for further computations.

- **Debiasing**: The sample_debiasing_t(shape) method performs non-uniform sampling of t to debias the weight standard deviation squared divided by the diffusion squared. This step is crucial for obtaining accurate estimates in certain scenarios and ensures the correctness of subsequent computations.

3.1.2 Multilayer Perceptron (MLP)

The MLP class represents a multilayer perceptron architecture used within the PluginReverseSDE class to model the drift function a of the reverse SDE. It consists of three hidden layers and applies the Swish activation function after each hidden layer.

The key elements of the MLP class are as follows:

- **Architecture**: The MLP has an input dimension, index dimension, and hidden dimension, which can be customized. It uses fully connected layers to process the input data and extract relevant features. The input dimension is denoted as $input_dim$, the index dimension as $index_dim$, and the hidden dimension as $hidden_dim$.

- **Activation Function**: The MLP employs the Swish activation function, which applies the sigmoid function to the input multiplied by the input itself. The Swish activation function is defined as $\text{Swish}(x) = \text{sigmoid}(x) \cdot x$.

3.1.3 Plugin Reverse SDE

The PluginReverseSDE class represents a reverse SDE obtained by inverting a base SDE. It utilizes the base SDE's drift function f and diffusion function g, along with an inference SDE's drift function a, to define the drift and diffusion functions of the reverse SDE.

The key elements of the PluginReverseSDE class are as follows:

- **Base SDE and Drift Function**: The base SDE represents the original SDE that is being inverted. The drift function of the base SDE, denoted as $f(t, y)$, is passed as an argument to the PluginReverseSDE class. The drift function of the reverse SDE is then defined as $\mu(t, y, \lambda) = (1 - 0.5\lambda) \cdot g(T - t, y) \cdot a(y, T - t) - f(T - t, y)$, where λ is an optional parameter that influences the drift.

- **Diffusion Function**: The diffusion function of the reverse SDE is defined as $\sigma(t, y, \lambda) = (1-\lambda)^{0.5} \cdot g(T - t, y)$, where λ again represents an optional parameter affecting the diffusion.

- **Denoising Score Matching Loss**: The PluginReverseSDE class provides a method called dsm(x) to compute the denoising score matching (DSM) loss. This loss function estimates the discrepancy between the true distribution and the estimated distribution of the data. It utilizes the sampled values from the base SDE and the drift function a to compute the loss.

- **Estimating the ELBO**: The elbo_random_t_slice(x) method estimates the evidence lower bound (ELBO) of the PluginReverseSDE. It samples t uniformly between 0 and T and employs the Hutchinson trace estimator to estimate div(mu). The ELBO is computed based on the sampled values, the drift function a, and additional terms related to the computation of the loss.

By combining the VariancePreservingSDE, MLP, and PluginReverseSDE classes, this implementation provides a comprehensive framework for working with variance-preserving SDEs and plugin reverse SDEs. It integrates mathematical formulations, sampling techniques, and neural network architectures to model and estimate parameters in a probabilistic setting.

3.1.4 Advantages of the framework in Generative Modeling

The proposed model, combining VariancePreservingSDE, MLP, and PluginReverseSDE, holds additional value in the context of generative models. Generative models aim to learn and generate new samples from a given distribution, capturing the underlying data patterns. The integration of SDEs and neural networks in this framework offers several advantages for generative modeling.

Firstly, the VariancePreservingSDE provides a flexible and efficient approach for sampling from the SDE. By preserving the variance of the underlying process, it ensures that the generated samples accurately represent the distribution. This is crucial for generative models, as it allows for the faithful replication of the data characteristics.

Secondly, the MLP architecture employed within the PluginReverseSDE enables the modeling of the drift function of the reverse SDE. The MLP's ability to learn complex nonlinear relationships makes it well-suited for capturing intricate patterns in the data. By leveraging the MLP, the PluginReverseSDE can effectively invert the base SDE and provide a generative model that produces samples from the desired distribution.

Furthermore, the denoising score matching loss implemented in the PluginReverseSDE adds an additional training objective for the generative model. By estimating the discrepancy between the true and estimated distributions, the model can further refine its performance and improve the quality of generated samples. This loss function contributes to the optimization process, encouraging the model to better capture the data distribution's characteristics.

Overall, the integration of variance-preserving SDEs, MLPs, and plugin reverse SDEs in this framework offers a comprehensive and powerful toolset for generative modeling. It combines the benefits of SDEs in accurately representing distributions, the flexibility of MLPs in modeling complex relationships, and the additional training objective of the denoising score matching loss. This integrated approach enhances the generative model's ability to capture intricate data patterns and generate high-quality samples.

3.2. Evaluation datasets:

To evaluate the SDE model, three distinct datasets were employed in the experiments: Moon dataset, S curve dataset, and Swiss roll dataset. Each dataset offers unique characteristics and poses different challenges, providing valuable insights into the model's performance in diverse scenarios, giving us insights into the added value of using the variational framework.

The Moon dataset 7, named after its crescent moon shape, consists of two intertwined half-moon shapes. This dataset is particularly useful for evaluating the model's ability to handle non-linear and complex patterns. By incorporating the Moon dataset, we aimed to assess the model's performance in capturing intricate relationships and making accurate predictions in a challenging setting.

The S curve dataset 6, as the name suggests, is shaped like the letter 'S'. It presents a curved pattern with two distinct parts that twist and merge. This dataset is commonly employed to evaluate models' capability to handle data with abrupt changes and sharp transitions. By including the S curve dataset in our experiments, we sought to investigate how well the SDE model could adapt to sudden variations and capture the underlying dynamics accurately.

The Swiss roll dataset 5 is a classic example used in machine learning for assessing models' ability to handle manifold learning and non-linear dimensionality reduction. It consists of a three-dimensional curved surface that resembles a rolled-up sheet of paper. By utilizing the Swiss roll dataset, our objective was to examine how effectively the SDE model could capture the underlying structure of the data and recover the original geometric shape.

By employing these three datasets, we aimed to comprehensively evaluate the SDE model's performance across various challenging scenarios. The Moon dataset assessed its capacity to handle non-linear patterns, the S curve dataset examined its adaptability to abrupt changes, and the Swiss roll dataset tested its ability to capture complex manifold structures. Collectively, these experiments provided valuable insights into the model's strengths, limitations, and potential for real-world applications.

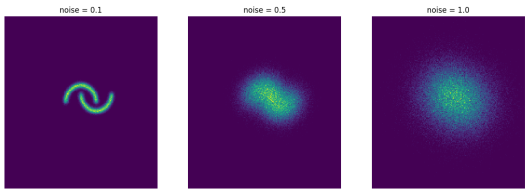

Figure 1. Example of Moon dataset

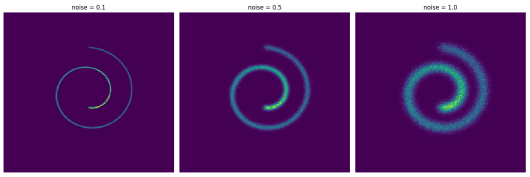

Figure 2. Example of Swiss roll dataset

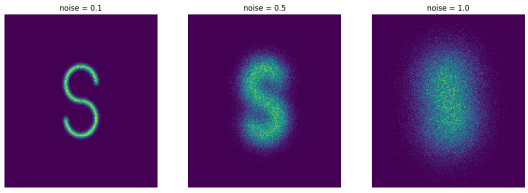

Figure 3. Example of S curve dataset

3.3. Results analysis:

In this study, we employed the proposed SDE model on three distint datasets: Moon dataset, Swiss Roll dataset, and Scurve datset. The model was trained separately on each dataset, and subsequently, we utilized the Euler Maruyama sampler to generate samples with varying lamda values. Our results revealed interesting observationsregarding the model's performance on these datasets.

Firstly, for the Swiss Roll and Scurve datasets, the trained model demonstrated promising outcomes in terms of sample quality. The generated samples exhibited characteristics that closely resembled the underlying data distributions, capturing the intricate patterns and structures present in the datasets. This success can be attributed to the ability of the model to effectively learn the drift and diffusion func-

tions, allowing it to accurately model the complex patterns in the training datasets.

However, the model encountered challenges when dealing with the Moon dataset. Despite the extensive training, the generated samples failed to capture the desired distribution adequately. This outcome suggests that the model struggled to capture the unique shape and multi-modal nature of the Moon dataset. The limitations could be attributed to the model's inherent assumptions or the complexity of the underlying data distribution itself, which may require a more tailored or specialized approach to accurately model its dynamics.

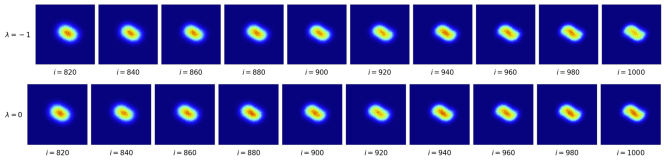

Figure 4. Samples from plug-in reverse SDEs of Moon dataset

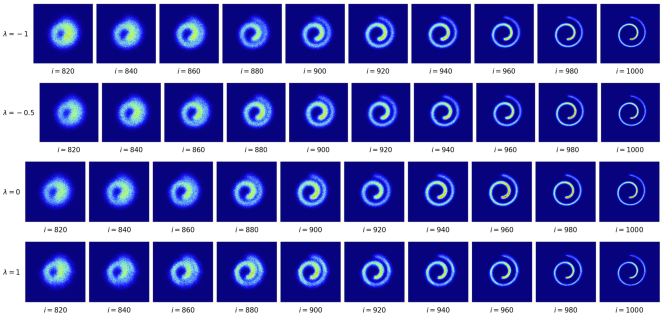

Figure 5. Samples from plug-in reverse SDEs of Swiss roll dataset

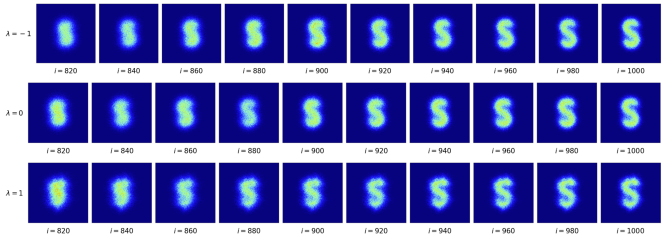

Figure 6. Samples from plug-in reverse SDEs of S curve dataset

Furthermore, an interesting observation during training was the existence of a noticeable gap between the score matching loss and the ELBO score. The score matching loss measures the discrepancy between the model's predicted score function and the empirical score function, while the ELBO (Evidence Lower Bound) serves as a measure of the model's overall performance in capturing the underlying data distribution. The presence of a gap between these two scores suggests a discrepancy between the quality of the individual samples generated by the model and the model's overall ability to capture the global characteristics of the data distribution. This observation implies that although the generated samples may exhibit local fidelity, the model may still lack the capability to fully capture the global structure of the dataset.

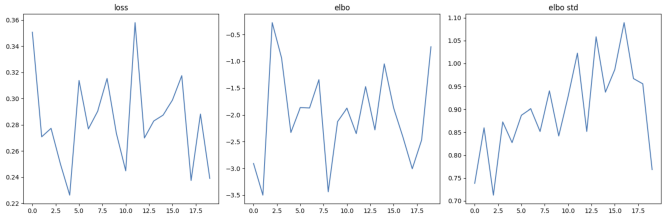

Figure 7. Loss and elbo distribution of Swiss Roll training

In summary, our experiments with the proposed model on the Moon, Swiss Roll, and S-curve datasets provided valuable insights into its performance. While the model demonstrated proficiency in capturing the complex distributions of Swiss Roll and S-curve datasets, it faced difficulties in accurately modeling the Moon dataset. The presence of a gap between the score matching loss and the ELBO score emphasizes the need for further investigation and refinement to ensure that the generated samples not only possess local fidelity but also capture the global characteristics of the underlying data distribution. These findings shed light on the model's strengths and limitations, paving the way for future research to enhance its performance and extend its applicability in generative modeling tasks.

4. Conclusion

In conclusion, the variational framework has brought a theoretical understanding of the score-matching loss by connecting it with maximum likelihood of the parametrized family of the plug-in reverse diffusion equations. Furthermore, from a practical standpoint, it has provided us with a specific weighting in the ISM loss to achieve maximum likelihood, which although not always beneficial is still often looked for in generative modeling.

On a broader note on generative modeling which is still in its early stages, we progressively observe that different models are connected that one could think at first glance. Normalizing flows can be seen as autoregressive models, and their continuous equivalent are deeply connected to diffusion models.

However it would now be interesting to look beyond likelihood. In [5], the authors advocate for the use of quantile loss functions instead of likelihoods because they allow for a less restrictive training on the parameters.

5. Authors contribution

Imane replicated the experiments on the new datasets and plotted the loss and elbo evolution over training (see

attached notebook). Nathan wrote the theoretical analysis of the results.

References

[1] Aaron Courville Chin-Wei Huang, Jae Hyun Lim. A variational perspective on diffusion-based generative models and score matching, 2021. 1, 2, 3

[2] Sohl-Dickstein J. Kingma D. P. Kumar A. Ermon S. Song, Y. and B. Poole. Score-based generative modeling through stochastic differential equations., 2021. 1, 2

[3] Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis, 2022. 3

[4] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models, 2021. 3

[5] Georg Ostrovski, Will Dabney, and Rémi Munos. Autoregressive quantile networks for generative modeling, 2018. 6