A Variational Perspective on Diffusion-Based Generative Models and Score Matching ?

Imane Elbacha, Nathan Weill

MVA 22-23

24/05/2023

Content

Motivation

- Providing a theoretical grounding in MLE of denoising diffusion probabilistic models (DDPM)

Main contributions

- Estimate the likelihood of continuous-time diffusion models
- Derive an efficient & scalable training method
- Proving the equivalence between maximizing the CT-ELBO and score matching

Likelihood of diffusion processes

Neural ODE & continuous normalizing flows

1. The flow is defined as the solution of
$\mathrm{d}X = \mu(X, t)\mathrm{d}t; \quad X(0) \sim p_0$
2. Its density is tractable with the instantaneous change of variable :
$$p_T(x) = p_0(Y(1))e^{\int_0^T -\nabla\cdot\mu(Y(s),T-s)ds}$$
$$\mathrm{d}Y = -\mu(Y, T-s)\mathrm{d}s; \quad Y(0) = x$$
3. Need to solve the ODE with numerical integration, not scalable

Likelihood of diffusion processes

Density of diffusion processes

1. Now take a stochastic diffusion process instead defined by : $\mathrm{d}X = \mu \mathrm{d}t + \sigma \mathrm{d}B_t$
2. E.g. Ornstein-Uhlenbeck or Geometric Brownian Motion
3. The density is not tractable anymore but the authors derive a continuous lower bound !

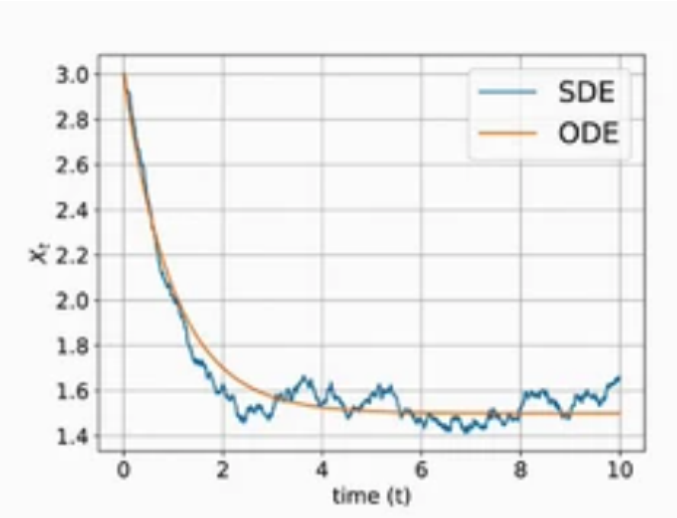


Figure – ODE vs SDE evolution

Likelihood of diffusion processes

CT-ELBO

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

  $X_0 \sim p(\cdot, 0)$
- Fokker-Planck PDE :

$$\partial_t p(x,t) = -\nabla \cdot (p(x,t)\mu(x,t)) + \frac{1}{2}\sigma^2(t)\Delta p(x,t)$$

$$= -(\nabla \cdot \mu)p - \mu^\top \nabla p + \frac{1}{2}\sigma^2 \Delta p$$

- And Feynman-Kac representation as an expectation :

$$p(x,T) = \mathbb{E}\left[p\left(Y_T, 0\right) e^{\int_0^T -\nabla \cdot \mu(Y_s, T-s)ds} \mid Y_0 = x\right]$$

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

$$Y_0 = x$$

Likelihood of diffusion processes

CT-ELBO (1)

- still intractable so variational inference

$$\log p(x,T) = \log \mathbb{E}_{Y\sim\mathbb{P}}\left[p(Y_T,0)e^{\int_0^T -\nabla\cdot\mu(Y_s,T-s)\,ds}\,\middle|\,Y_0=x\right]$$

Marginalization

$$= \log \mathbb{E}_{Y\sim\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}}\,p(Y_T,0)e^{\int_0^T -\nabla\cdot\mu(Y_s,T-s)\,ds}\,\middle|\,Y_0=x\right]$$

Change of measure

$$\geq \mathbb{E}_{Y\sim\mathbb{Q}}\left[\log\frac{d\mathbb{P}}{d\mathbb{Q}} + \log p(Y_T,0) - \int_0^T \nabla\cdot\mu\,ds\,\middle|\,Y_0=x\right]$$

Jensen's inequality

KL divergence

Likelihood of diffusion processes

CT-ELBO (2)

- Using Girsanov theorem

$$\left.\begin{aligned}\mathbb{P}\text{: } dY &= -\mu ds + \sigma dB'_s \\ \mathbb{Q}\text{: } dY &= (-\mu + \sigma a)ds + \sigma d\hat{B}_s\end{aligned}\right\} \Rightarrow D_{\mathrm{KL}}(\mathbb{Q}\,||\,\mathbb{P}) := \mathbb{E}_{\mathbb{Q}}\left[\log \frac{d\mathbb{Q}}{d\mathbb{P}}\right] = \mathbb{E}_{\mathbb{Q}}\left[\int_0^T \|a(Y_s, s)\|^2 ds\right]$$

Inference SDE

$$\log p(x, T) \geq \mathbb{E}_{Y\sim\mathbb{Q}}\left[\log \frac{d\mathbb{P}}{d\mathbb{Q}} + \log p(Y_T, 0) - \int_0^T \nabla \cdot \mu \, ds \,\middle|\, Y_0 = x\right]$$

$$= \mathbb{E}_{Y\sim\mathbb{Q}}\left[\log p(Y_T, 0) - \int_0^T \nabla \cdot \mu + \|a\|^2 ds \,\middle|\, Y_0 = x\right]$$

25

Figure – Girsanov theorem

Likelihood of diffusion processes

CT-ELBO (3)

- MCMC sampling

$$\mathscr{E}^{\infty} := \mathbb{E}_{Y\sim\mathbb{Q}}\left[\log p(Y_T,0) - \int_0^T \nabla\cdot\mu + \|a\|^2\, dr \,\middle|\, Y_0 = x\right]$$

$$= \mathbb{E}_{Y_T}\left[\log p(Y_T,0)\right] - \int_0^T \mathbb{E}_{Y_s}\left[\|a\|^2 + \nabla\cdot(ga - f) \,|\, Y_0 = x\right] \mathrm{d}s$$

No need to integrate if we can sample

Figure – MCMC sampling

Likelihood of diffusion processes

CT-ELBO (4)

- Noisier but faster than numerical integration

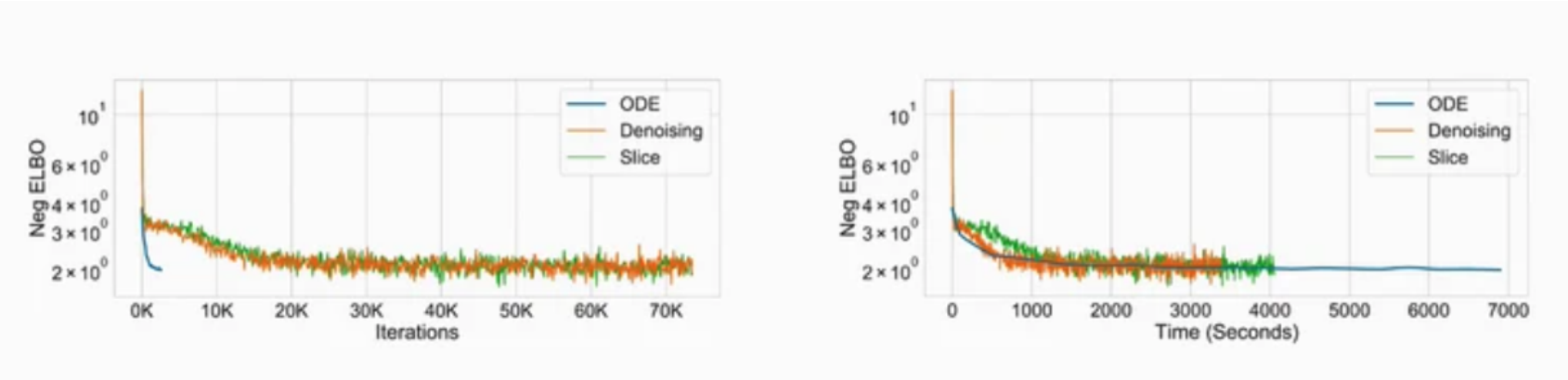


Figure – Negative ELBO evolution over training

Connection to score matching

The reverse plug-in SDE

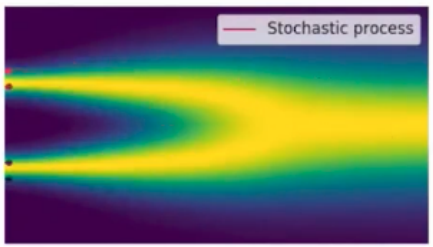
- score matching with specific weighting and with ISM loss is found to be equivalent to maximizing the CT-ELBO of the reverse plug-in SDE

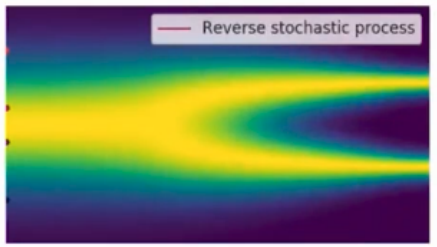
(setting $a = g\mathbf{s}_\theta$)

$$-\mathbb{E}_{Y_0}[\mathscr{E}^\infty] \overset{+\text{cst}}{=} \int_0^T g^2 \mathbb{E}_{Y_s}\left[\|\mathbf{s}_\theta\|^2 + \nabla \cdot \mathbf{s}_\theta\right] \mathrm{d}s \qquad \text{(Implicit score matching)}$$

$$\overset{+\text{cst}}{=} \int_0^T g^2 \mathbb{E}_{Y_s}\left[\|\mathbf{s}_\theta(Y_s, s) - \nabla \log q(Y_s, s)\|^2\right] \mathrm{d}s \qquad \text{(Explicit score matching)}$$ Hyvärinen 05'

$$\mathrm{d}Y = f\mathrm{d}s + g\mathrm{d}\hat{B}_s \qquad \mathrm{d}X = (g^2 \overbrace{\nabla \log q(y,s)}^{\approx \mathbf{s}_\theta(y,s)} - f)\mathrm{d}t + g\mathrm{d}B_t$$

Song 21'

Connection to score matching

Parametrized SDE family

$$\mathrm{d}X = \left(\left(1 - \frac{\lambda}{2}\right) g^2 \overbrace{\nabla \log q}^{\approx \mathbf{s}_\theta(y,s)} - f \right) \mathrm{d}t + \sqrt{1-\lambda} g \, \mathrm{d}B_t \qquad \mathrm{d}Y = \left(f - \frac{\lambda}{2} g^2 \nabla \log q \right) \mathrm{d}s + \sqrt{1-\lambda} g \, \mathrm{d}\hat{B}_s$$

- $\lambda = 0$ corresponds to the previous SDE
- $\lambda = 1$ corresponds to the ODE
- Connection to what we saw in class (same marginal densities but different dynamics)

- As a consequence the two following dynamics have the **same** marginal densities.
  - $\mathrm{d}\mathbf{X}_t = b(t, \mathbf{X}_t)\mathrm{d}t + c^{1/2}\mathrm{d}\mathbf{B}_t$
  - $\mathrm{d}\mathbf{X}_t = \{b(t, \mathbf{X}_t) - \frac{c}{2}\nabla \log p_t(\mathbf{X}_t)\}\mathrm{d}t.$
  - One is **deterministic**, the other is **stochastic**.

Connection to score matching

Link with discrete ELBO

- we saw in class variational inference in discrete formulation
- arose a connection with the DSM loss
- Similar result found in **?**

Experiments

Overview

1. Implement reverse SDE training on three datasets : moon dataset, Swiss roll dataset, and S-curve dataset.
2. Use the pluging reverse SDE training to create respective models for each dataset.
3. Perform sampling on the generated models using Euler-Mayumar method.

Experiments

Implementation of pluging reverse SDE

- VariancePreservingSDE : Implements a variance-preserving stochastic differential equation (SDE) with key parameters, drift and diffusion functions, and sampling methods.
- MLP : Represents a customizable multilayer perceptron architecture with Swish activation function, used for modeling the drift function in PluginReverseSDE.
- PluginReverseSDE : Defines a reverse SDE by inverting a base SDE, utilizing drift and diffusion functions from both SDEs. Provides methods for denoising score matching loss and estimating the evidence lower bound (ELBO).

Experiments

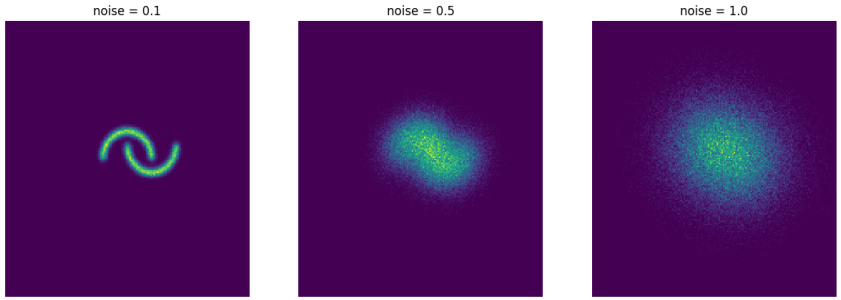
Advantages of pluging reverse SDE

- The VariancePreservingSDE ensures accurate representation of the distribution by preserving the variance of the underlying process during sampling.
- The MLP architecture within the PluginReverseSDE effectively models complex patterns, enabling the generation of samples from the desired distribution.
- The denoising score matching loss in the PluginReverseSDE further refines the generative model by estimating the discrepancy between true and estimated distributions, improving the quality of generated samples.

Experiments

Datasets

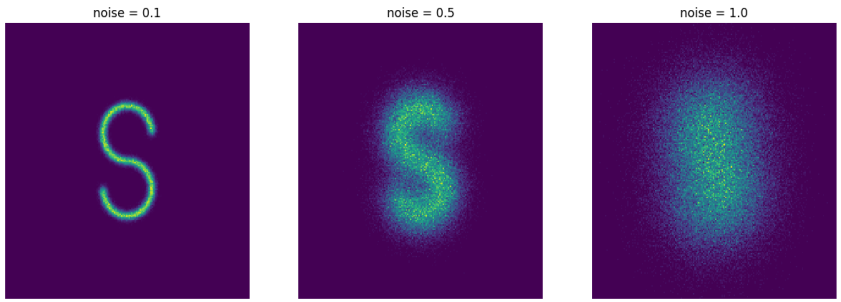
- Moon dataset : Used to evaluate the model's performance in handling non-linear and complex patterns, providing insights into its ability to capture intricate relationships and make accurate predictions in challenging scenarios.



Experiments

Datasets

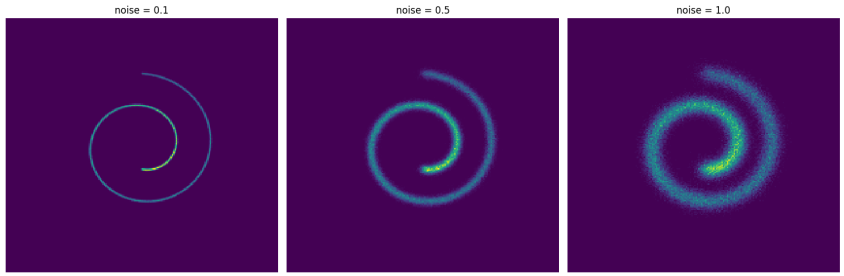
- S curve dataset : Employed to assess the model's capability to handle data with abrupt changes and sharp transitions, examining its adaptability to sudden variations and its capacity to capture the underlying dynamics accurately.



Experiments

Datasets

- Swiss roll dataset : Utilized to evaluate the model's ability to handle manifold learning and non-linear dimensionality reduction, examining its effectiveness in capturing the underlying structure of the data and recovering the original geometric shape.



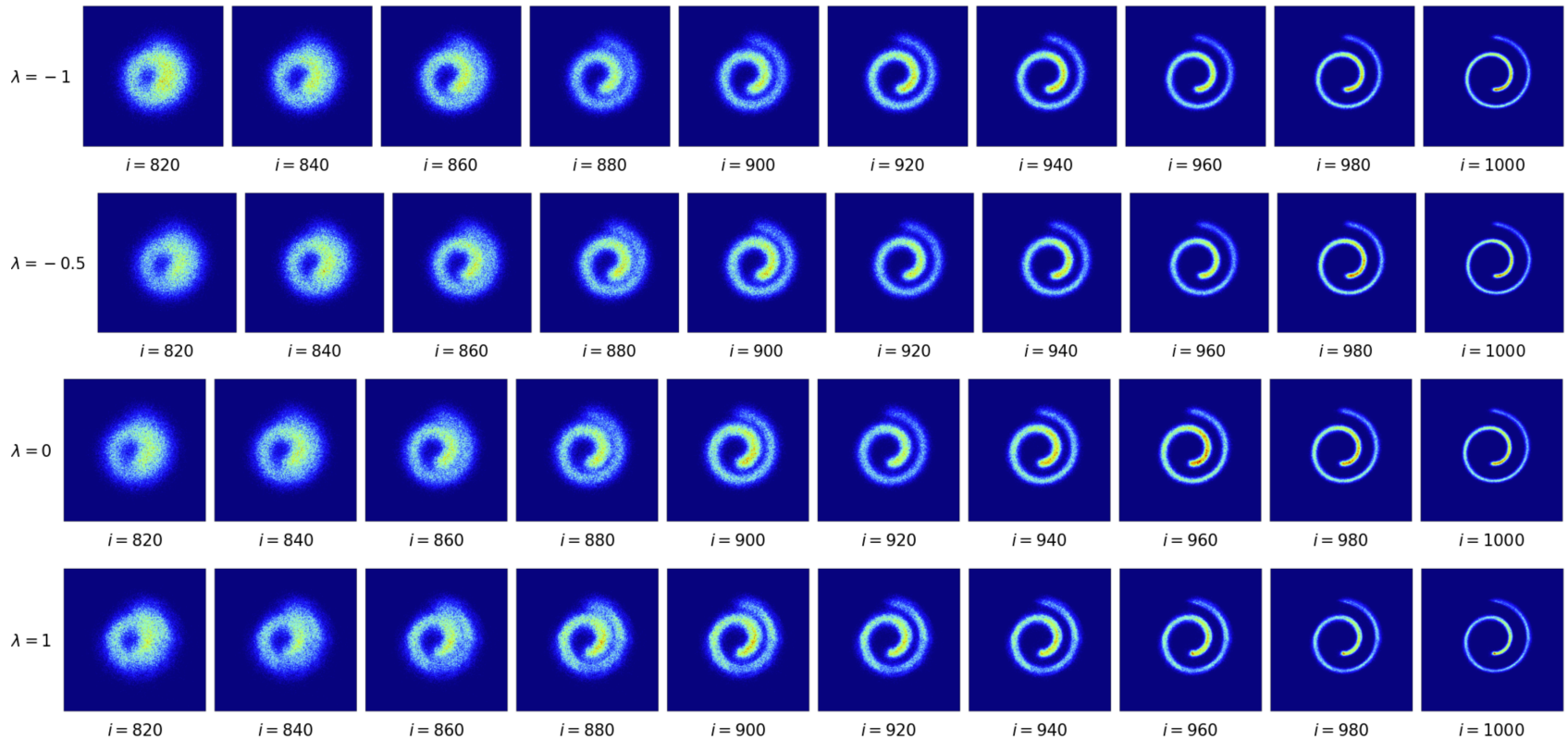
Experiments

Sampling results

- The proposed SDE model was applied to three distinct datasets : Moon dataset, Swiss Roll dataset, and S-curve dataset.
- The model demonstrated promising results for the Swiss Roll and S-curve datasets, accurately capturing the patterns and structures present in the data.
- However, the model encountered difficulties in generating samples that accurately represented the distribution of the Moon dataset, indicating challenges in capturing its unique shape and multi-modal nature.

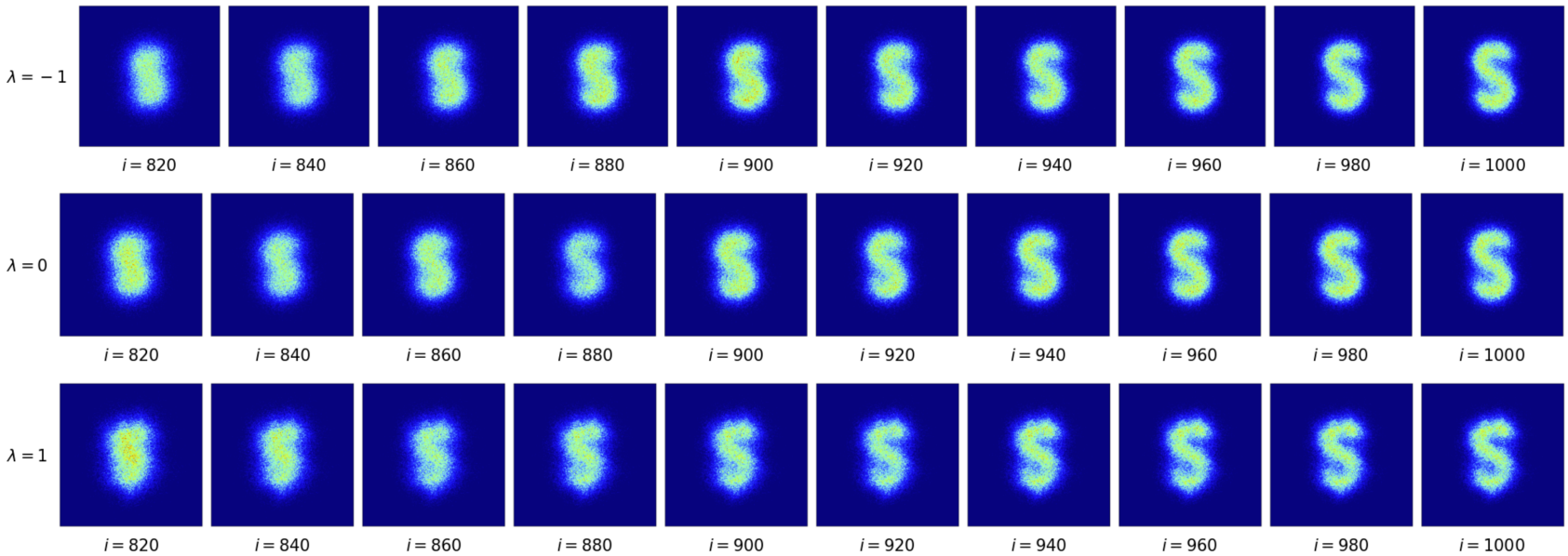
Experiments

Sampling results



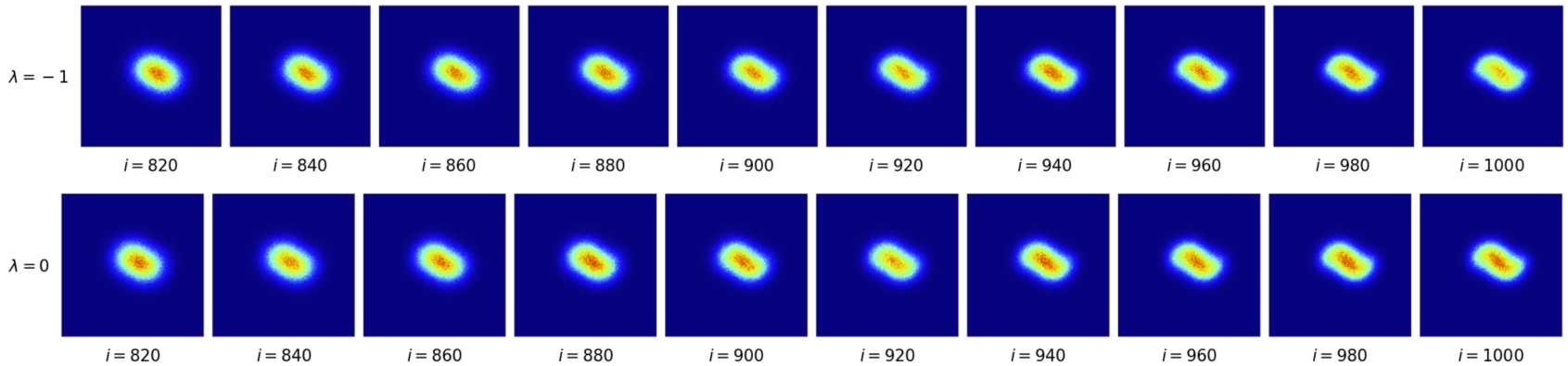
Experiments

Sampling results



Experiments

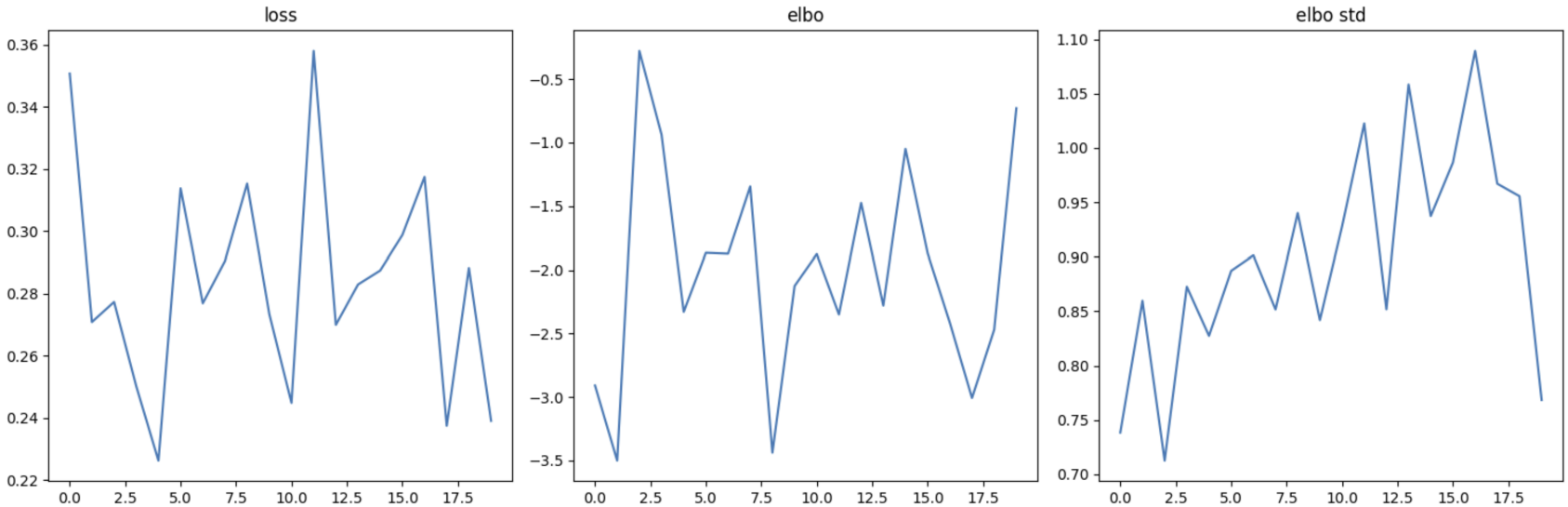
Sampling results



Experiments

Loss and Elbo discrepancy

- A gap was observed between the score matching loss and the ELBO score during training, indicating a discrepancy between the quality of individual samples and the model's ability to capture the global characteristics of the data distribution.



Conclusion

What is the importance of a variational framework ?

- The variational framework connects the score-matching loss to maximum likelihood estimation in the plug-in reverse diffusion equations.
- It provides a specific weighting in the importance sampling (ISM) loss for achieving maximum likelihood in generative modeling.
- Different generative models, such as normalizing flows and diffusion models, are interconnected.
- Beyond likelihood, alternative loss functions like quantile loss offer less restrictive parameter training, as suggested by Ostrovski et al. (2018).

References I