Hessian Geometry of Latent Space in Generative Models

Alexander Lobashev, Dmitry Guskov, Maria Larchenko, Mikhail Tamm

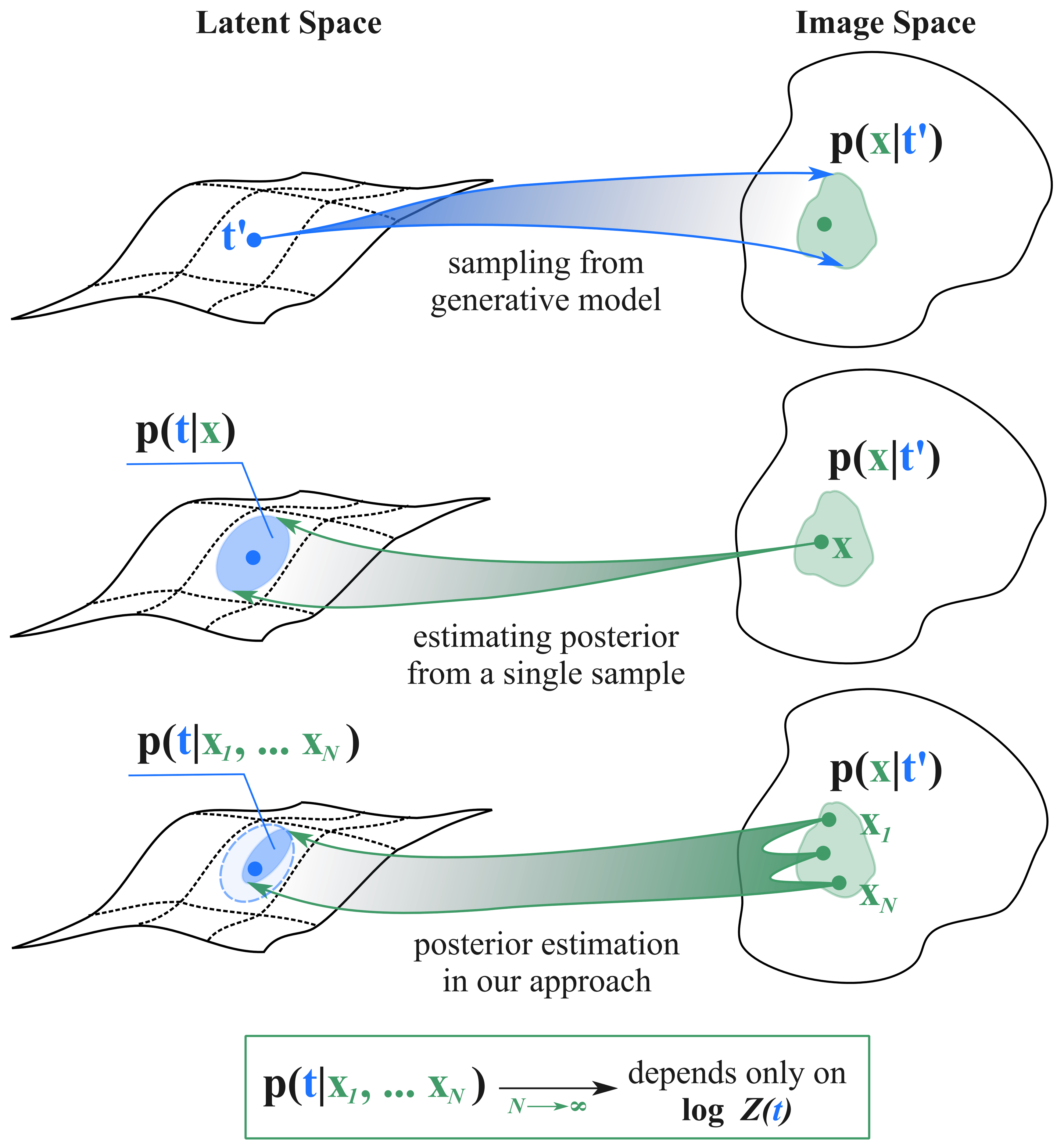
Glam AI • Artificial Neural Computing Corp. • Magicly AI • Tallinn University

Outline

Generative models often exhibit abrupt, non-smooth changes during latent space interpolation. We propose a novel method to analyze this phenomenon by unifying concepts from information geometry and statistical physics to map the geometric structure of the latent space.

- **Our Goal:** Reconstruct the Fisher Information Metric on the latent space to understand its geometry, identify phase transitions, and compute smoother interpolation paths (geodesics).
- **Our Method:** We approximate the posterior distribution of latent variables given generated samples. This allows us to learn the log-partition function $\log Z(t)$, whose Hessian defines the Fisher metric.
- **Theoretical justification** of our method is given by Theorem 3.1

Core Method



Theoretical Foundation (Thm. 3.1): The posterior distribution over latent parameters t concentrates around the true parameter t', with a shape defined by the Bregman divergence of the log-partition function $\log Z(t)$.

$$\lim_{N\to\infty} (p(t|x_1,\dots,x_N))^{1/N} \propto e^{-D_{\log Z(t)}(t,t')}$$

Two-Step Workflow:

1. Approximate Posterior $p(t|x)$:
   - For Physics Models: Train a U-Net on microstates (e.g., spin configurations).
   - For Diffusion Models: Use a pre-trained feature extractor (CLIP) to define a posterior based on embedding distances.
2. Learn the Metric from $\log Z(t)$:
   - Model $\log Z_\theta(t)$ with a neural network (MLP).
   - Train by minimizing the Jensen-Shannon Divergence between the approximated posterior and the model's derived posterior.
   - The Fisher metric is the Hessian of the learned function: $g_F = \nabla^2 \log Z_\theta(t)$.

Table: Quantitative Comparison for Diffusion Models.

Metric	Geodesic (Ours)	Linear	Geodesic (Shao/Wang)
CLIP Length	**72.3 ± 4.0**	73.6 ± 3.5	73.6 ± 4.4
Perceptual Path Length	**3.12 ± 0.16**	3.17 ± 0.23	3.19 ± 0.21
Mean Curvature	**0.37 ± 0.69**	0.00 ± 0.00	1.33 ± 0.53

Fractal Phase Boundaries in Diffusion Models

Applying our method to a 2D slice of Stable Diffusion's latent space reveals a complex phase diagram. The boundaries between distinct concepts (e.g., "lion" vs. "mountain") are not simple lines but exhibit a self-similar, fractal structure.

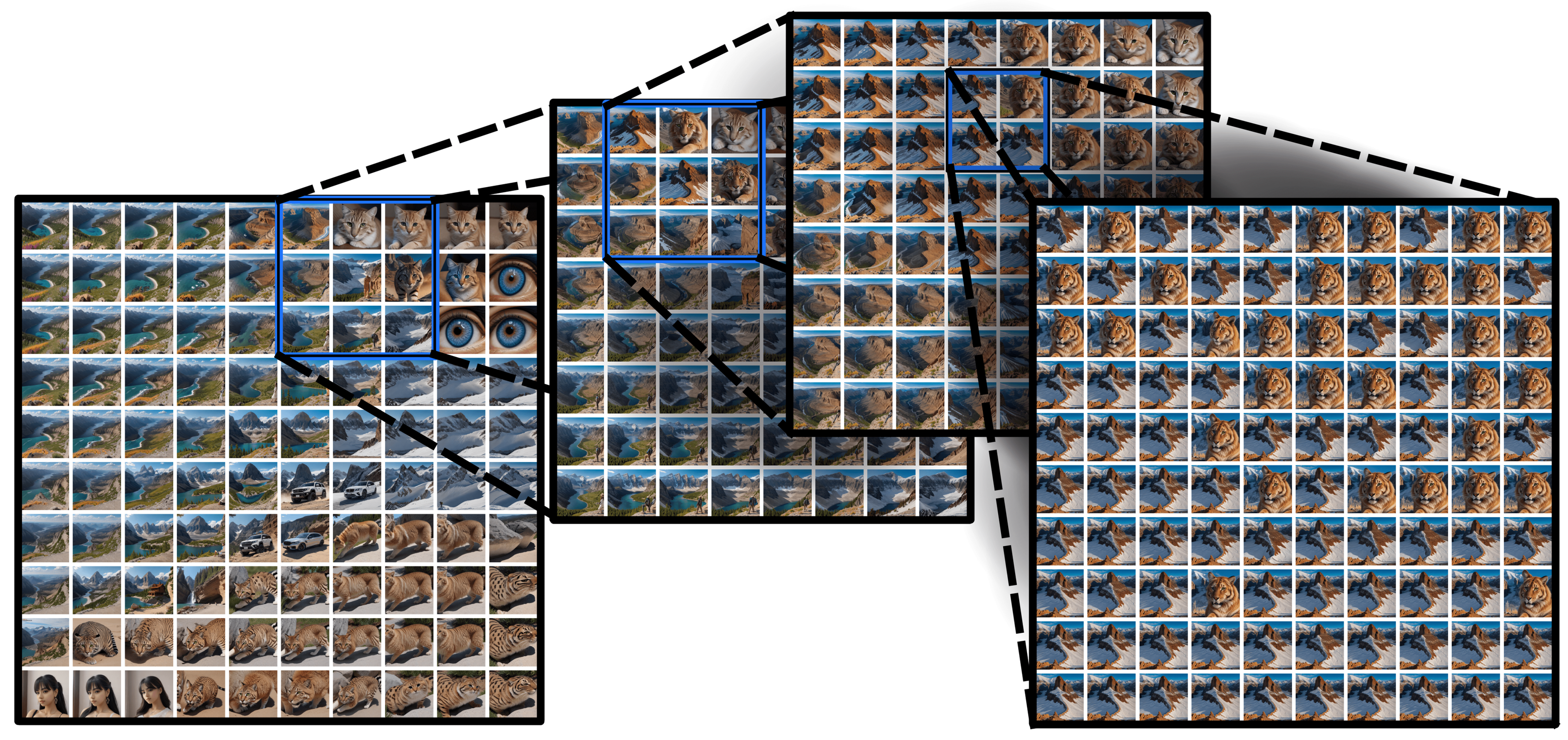


Figure: A fractal phase boundary. Zooming in reveals that the "lion" and "mountain" phases permeate each other at increasingly fine scales. The bottom-right plot shows a latent space variation of only 10^{-5} between adjacent images.

Latent Space Geometry & Geodesics

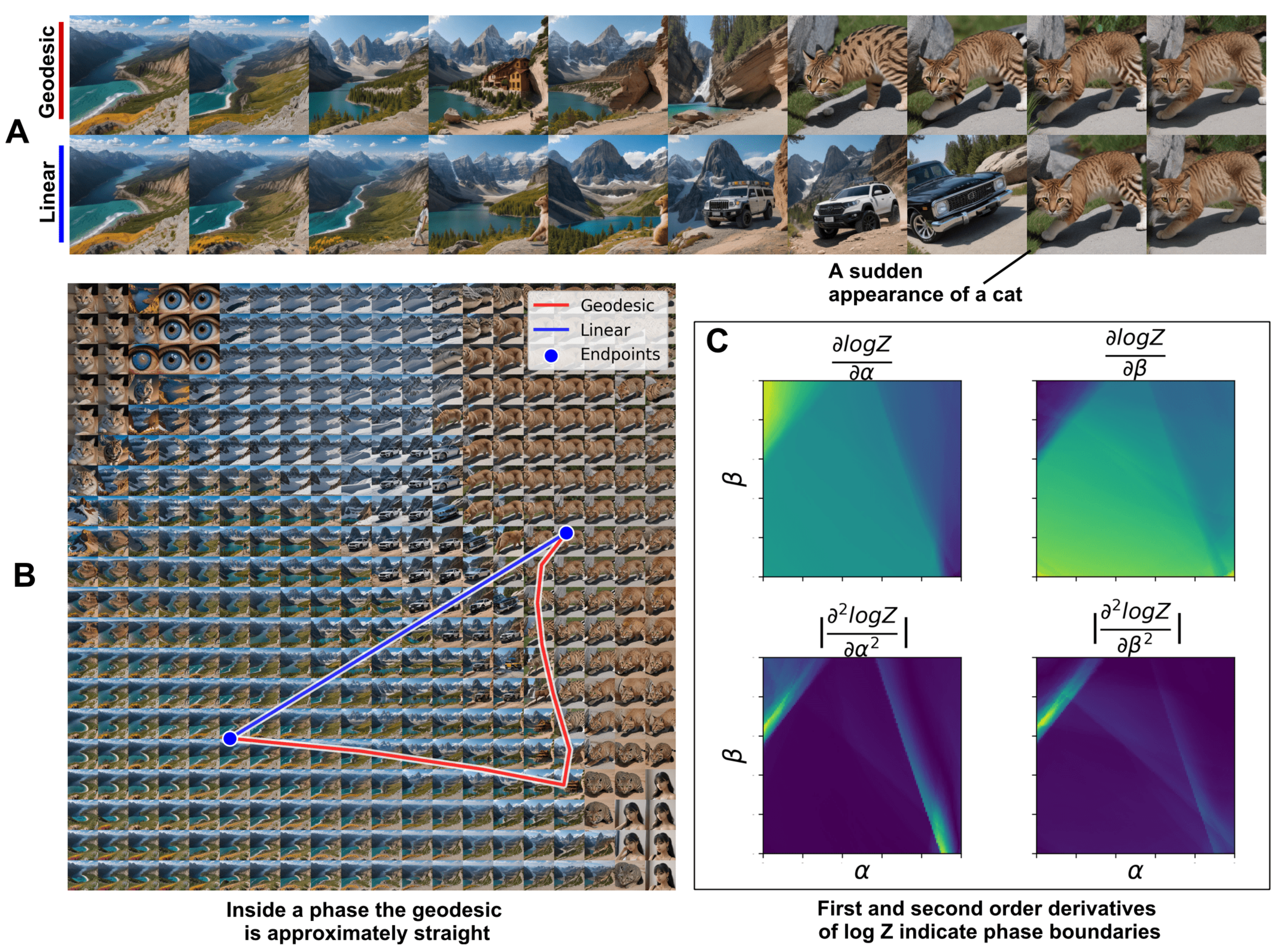


Figure: **(A)** Geodesic interpolation (top) is perceptually smoother than linear interpolation (bottom). **(B)** The latent space phase map shows distinct regions. Geodesics (red/blue paths) curve to navigate this geometry. **(C)** Metric components ($\partial^2 \log Z/\partial t_i \partial t_j$) show sharp peaks and discontinuities precisely at phase boundaries.

Underlying Mechanism For Phase Transition

Diverging Lyapunov exponent (Proposition 4.1)

Suppose that the (target) data distribution is a bimodal mixture of two Gaussians, each with variance σ^2:

$$p_0(x) = \frac{1}{2}\mathcal{N}(x \mid -1, \sigma^2) + \frac{1}{2}\mathcal{N}(x \mid 1, \sigma^2).$$

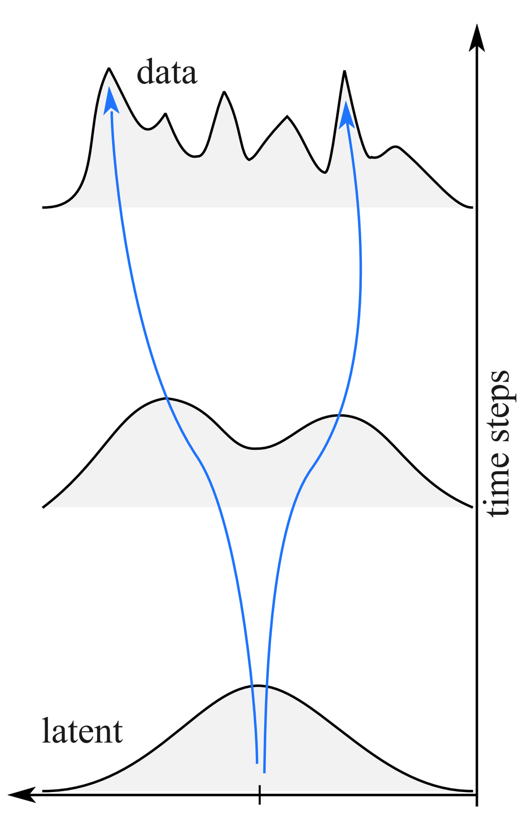
The latent distribution is the standard normal $\mathcal{N}(x \mid 0, 1)$. Consider the variance-preserving SDE

$$dX_t = -\frac{1}{2}\beta X_t\,dt + \sqrt{\beta}\,dW_t.$$

Then the Lyapunov exponent of the corresponding reverse-time ODE at $x = 0$ has the following form:

$$\lambda = \frac{\beta}{2}\left(1 + \frac{1-\sigma^2}{\sigma^4}\right),$$

and it diverges to infinity as $\sigma \to 0$. Then the point $x = 0$ can be seen as a phase transition boundary.



Contacts & Code

Alexander Lobashev
lobashevalexander@gmail.com

Dmitry Guskov
guskov01dmitry@gmail.com

Code

arXiv