Denoising Score Matching

Victorin Turnel, Thomas Winninger

Master MVA - ENS Paris-Saclay

Motivation

Modern generative models (Stable Diffusion, DALL-E) rely on **score-based denoising**, which learns gradients instead of densities, but why?

The Partition Function Problem:

Energy-based models define: $p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z_\theta}$

- $E_{\theta(x)}$: Energy function (the neural network).
- $Z_\theta = \int \exp\left(-E_{\theta(x)}\right) \mathrm{d}x$: Normalization constant.

Issue: For high-dim images ($d \approx 10^6$), computing Z_θ is **intractable**.

2. The Score Matching Solution : By modeling the gradient of the log-density (the score), Z_θ vanishes!

$$\psi_{\theta(x)} = \nabla_x \log p_{\theta(x)} = -\nabla_x E_{\theta(x)}$$

Z_θ is eliminated because it does not depend on x.

3. Manifold Hypothesis : Real data resides on low-dimensional manifolds. The score is undefined in empty space → Therefore, the solution is to perturb data with noise (NCSN).

Score Matching Framework

Goal: Learn the score function $s_\theta(x) \approx \nabla_x \log p_{\text{data}}(x)$ to bypass the intractable partition constant Z_θ.

1. Implicit Score Matching (ISM) (Hyvärinen, 2005) Minimizes the Fisher divergence with real data:

$$J_{\text{ISM}(\theta)} = \mathbb{E}_{p_{\text{data}}}\left[\frac{1}{2}\|s_\theta(x)\|^2 + \operatorname{tr}(\nabla_x s_\theta(x))\right]$$

→ **Problem:** Even though no partition function and no true score are needed, computing the Jacobian trace is $\mathcal{O}(d^2)$, which is intractable for high-dimensional images.

2. Denoising Score Matching (DSM) (Vincent, 2011) Perturb data with noise $\tilde{x} = x + \sigma\varepsilon$, then match the **conditional** score:

$$J_{\text{DSM}(\theta)} = \mathbb{E}_{q_\sigma(\tilde{x}|x)}\left[\frac{1}{2}\left\|s_\theta(\tilde{x}) - \underbrace{\frac{x - \tilde{x}}{\sigma^2}}_{\text{Target Score}}\right\|^2\right]$$

→ **Key Insight:** Now this alternate objective, inspired by denoising autoencoders, is equivalent to explicit score matching. No Hessian trace needed!

3. Noise Conditional Score Networks (NCSN) (Song and Ermon, 2020)

→ **Issue:** The score is undefined in low-density regions (Manifold Hypothesis)

→ **Solution:** Train a single network $s_\theta(x, \sigma)$ conditioned on geometric noise levels $\sigma_1 > \ldots > \sigma_L$ to populate the ambient space.

Sampling: Annealed Langevin Dynamics

Once the score $s_\theta(x, \sigma)$ is learned, how do we generate images?

1. Standard Langevin Dynamics Start from random noise x_0 and iteratively follow the score gradients towards high-density regions:

$$x_{t+1} = x_t + \frac{\varepsilon}{2} s_\theta(x_t) + \sqrt{\varepsilon} z_t, \quad z_t \sim \mathcal{N}(0, I)$$

→ **Limitation:** Fails to cross low-density regions between modes (poor mixing).

2. Annealed Dynamics (The Fix) (Song and Ermon, 2020) Use the learned noise levels $\sigma_1 > \ldots > \sigma_L$ as a schedule:

- **Start (High σ):** Large steps explore the whole space (good mixing).
- **End (Low σ):** Small steps refine details on the data manifold.

Algorithm: For each noise level σ_i:

$$x_{t+1} \leftarrow x_t + \frac{\alpha_i}{2} s_\theta(x_t, \sigma_i) + \sqrt{\alpha_i} z_t$$

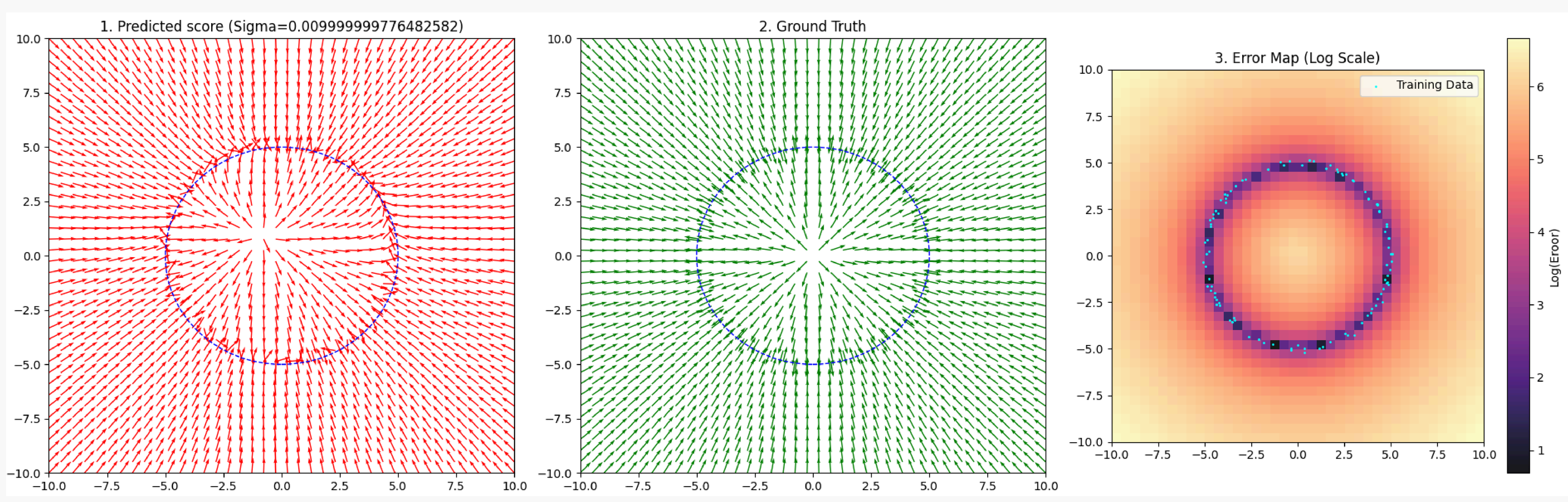
where step size α_i decreases with σ_i.

Experiments: Toy Data & Intuition

Before generating images, we validate the method on 2D toy distributions.

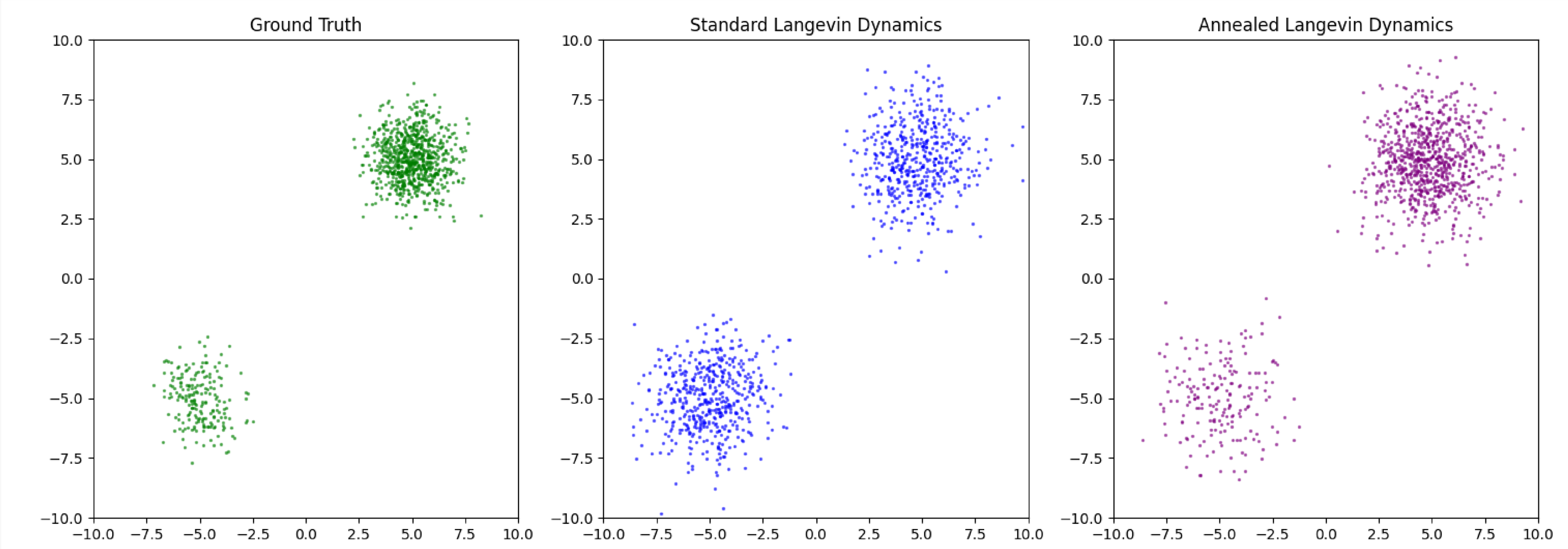
1. Visualizing the Score Field We trained a simple MLP on a "Circle" distribution.

- The learned score $s_{\theta(x)}$ forms a vector field pointing towards the high-density manifold.
- **Observation:** The score is accurate near data but undefined/random far from it.



2. Solving the Mixing Problem Task: Sample from a Mixture of Gaussians : $p_{\text{data}} = \frac{1}{5}\mathcal{N}((5,5), I) + \frac{4}{5}\mathcal{N}((-5,-5), I)$.

- **Standard Langevin:** Gets stuck in one mode; fails to recover the distribution weights.
- **Annealed Dynamics:** Large noise steps allow the chain to cross low-density regions and recover both modes correctly (see Fig. 2).



→ **Takeaway:** Multi-scale noise is mandatory for multimodal data!

Image Generation & Stability Analysis

Moving to real images (MNIST), we implemented a simplified U-Net from scratch.

1. The Instability Problem Standard training exhibits high variance.

- **Observation:** Note the huge spike at Epoch 10 (FID ≈ 11.3) for the Custom model.
- **Cause:** The score network oscillates around the manifold.

2. The Solution: EMA Exponential Moving Average ($\theta' \leftarrow m\theta' + (1-m)\theta_i$) stabilizes weights.

- **Result:** FID drops consistently to 0.22.

3. Hyperparameters Optimal sampling requires small step size $\varepsilon \approx 10^{-5}$ and large $T = 100$ to avoid "overshooting" (snow noise).

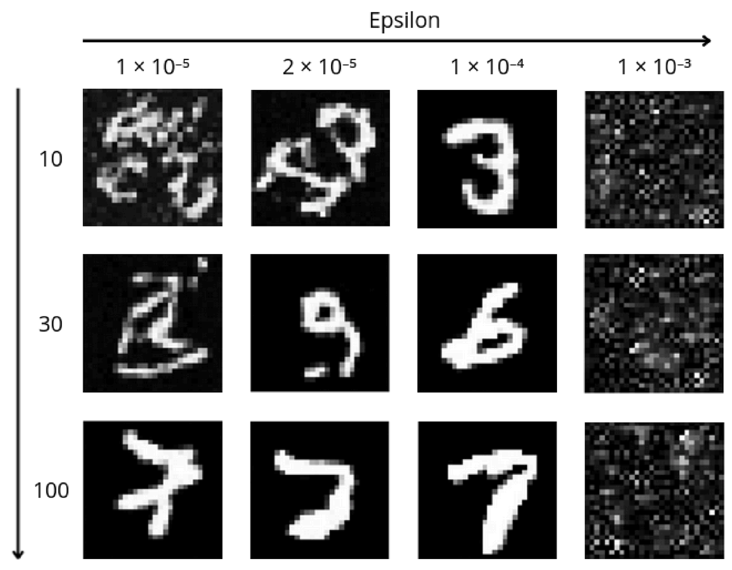


Figure 1: Impact of ε and T on sampling

Epoch	FID (Cust) ↓	FID (EMA) ↓	Loss (Cust)	Loss (EMA)
1	-	-	0.4502	0.3068
5	2.9746	0.6112	0.2095	0.1288
10	11.3181	0.3198	0.1609	0.1058
15	0.9717	0.2249	0.1471	0.0971

Table 1: Impact of EMA on Stability (FID scores from Report Table 1)

Using the U-Net architecture from (Song and Ermon, 2020) we tested training the model on a different dataset: Fashion MNIST.

4. Model Collapsing Small models tend to overfit one class and forget the others.

- **Observation:** Our custom U-Net only learned the "shirt" class, while the U-Net from (Song and Ermon, 2020) barely learned other classes at the beginning of the training before collapsing.
- **Cause:** One potential problem could be the capacity of the model, so we tested increasing the dimensions and adding dropout, which helped maintaining stability for a longer time. Experiments with other parameters (σ, T, ε) did not improve stability.

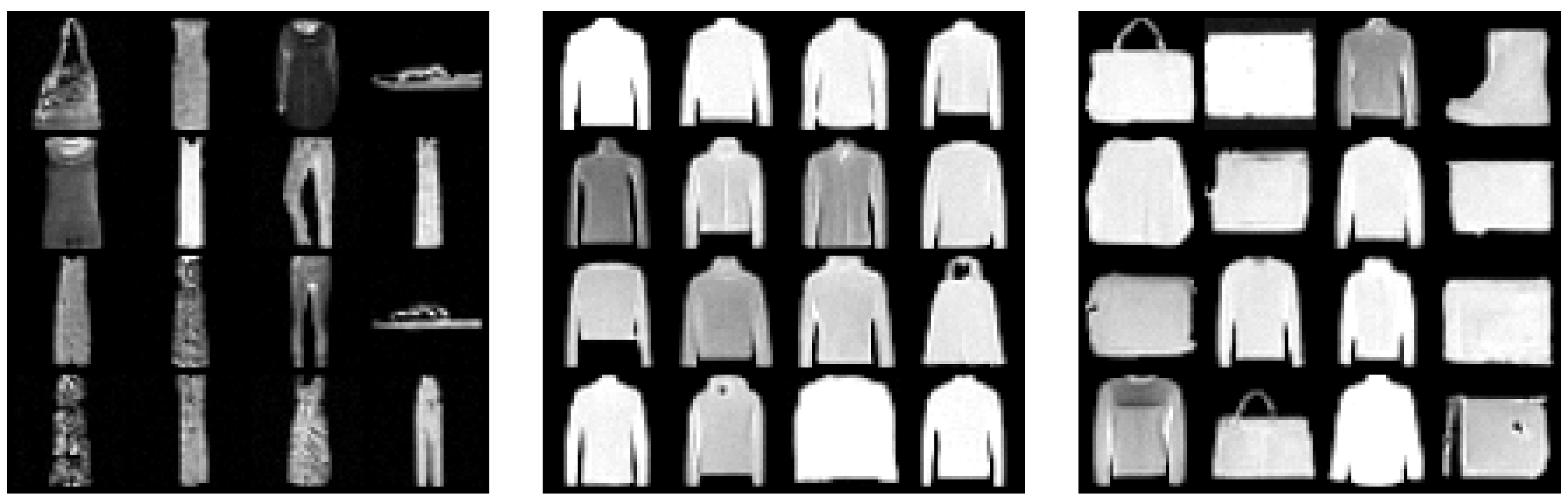


Figure 2: **Left and middle:** (Song and Ermon, 2020)'s U-Net on Fashion MNIST sampled at 30000 and 40000 epochs. **Right:** Slightly bigger model with dropout, sampled at 40000 epochs.

Future directions

Several directions merit exploration:

- ODE samplers can be improved;
- Diffusion can be used in other applications, like audio, 3D shapes, or molecules;
- Theoretical work can be done to understand why diffusion generalizes so well.

References

Hyvärinen, A. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24), 695–709, 2005.

Song, Y., and Ermon, S. *Generative Modeling by Estimating Gradients of the Data Distribution*, 2020.

Vincent, P. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7), 1661–1674, 2011. https://doi.org/10.1162/NECO_a_00142

victorin.turnel@ens-paris-saclay.fr, thomas.winninger@telecom-sudparis.eu