How Much Noise Is Enough? Exploring ϵ in Stochastic Interpolants

MATHÉMATIQUES VISION APPRENTISSAGE

Probabilistic Graphical Models — MVA 2025–2026

Janis Aiad Guillaume Bernard Félix Rosseeuw

école normale supérieure paris—saclay

Introduction

Score-based generative models turn noise into data by following a gradient field (Langevin / diffusion). *Stochastic interpolants* generalize this idea by separating deterministic transport from stochasticity. **Our experimental question is simple: how much noise ϵ do we really need?** On CelebA, we explore several paths $I(t,\cdot,\cdot)$, $\gamma(t)$ functions, and ϵ values, and show how ϵ controls the trade-off between *fidelity* (preserving content) and *robustness/diversity* (exploring without collapsing).

DAEs and Score Matching

Denoising autoencoders. Given a corrupted data

$$\tilde{x} = x + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

a *DAE* r_θ is trained by minimizing

$$J_{\mathrm{DAE}_\sigma}(\theta) = \mathbb{E}_{q_\sigma(\tilde{x},x)}\left[\|x - r_\theta(\tilde{x})\|^2\right].$$

Score matching. Instead of modeling a density $q(x)$, s_θ can approximate its *score* $\nabla_x \log q(x)$, but the ideal objective

$$J_{\mathrm{ESM}_q}(\theta) = \mathbb{E}_{q(x)}\left[\tfrac{1}{2}\|s_\theta(x) - \nabla_x \log q(x)\|^2\right]$$

is not directly usable because the score of q is unknown.

Denoising score matching. Given a Parzen density estimator q_σ, Vincent [?] introduces *denoising score matching*

$$J_{\mathrm{DSM}_{q_\sigma}}(\theta) = \mathbb{E}_{q_\sigma(\tilde{x},x)}\left[\frac{1}{2}\left\|s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x} \mid x)\right\|^2\right],$$

where the target score $\nabla_{\tilde{x}} \log q_\sigma(\tilde{x} \mid x)$ is shown proportional to $x - \tilde{x}$, the denoising direction. Vincent shows that

$$J_{\mathrm{DAE}_\sigma} \smile J_{\mathrm{DSM}_{q_\sigma}} \smile J_{\mathrm{ESM}_{q_\sigma}},$$

so training a DAE is (up to constants) equivalent to learning a score field.

Implication. Denoising $\approx$ score learning.

References

[1] Hyvärinen (2005) Score Matching. [2] Vincent et al. (2008) Denoising Autoencoders.
[3] Vincent (2011) Connection between score matching and DAEs. [4] Song & Ermon (2019) Score-based generative modeling.
[5] Albergo et al. (2023) Stochastic Interpolants.

Score-Based Generative Modeling

Key idea. Instead of learning $q(x)$, learn its score $\nabla_x \log q(x)$. This vector field tells, locally, which direction increases probability, so it can drive Langevin sampling dynamics from a random initialization to data.

Why naive Langevin breaks down. Natural data are often assumed to lie near a low-dimensional manifold embedded in the ambiant space (*manifold hypothesis*). This creates two fundamental issues for score learning: (i) the score $\nabla_x \log q(x)$ is ill-defined off the data support, and (ii) standard score matching objectives become inconsistent when the data do not cover the ambient space.

Even when the manifold-related issues are addressed, Langevin dynamics can still struggle in practice. Large low-density regions—where the data provide little to no supervision—make the score unreliable away from the data. And even with a reasonably accurate score, Langevin may still mix slowly: trajectories can linger in low-density areas (Figure 1), and samples can allocate samples in the wrong proportions across modes.

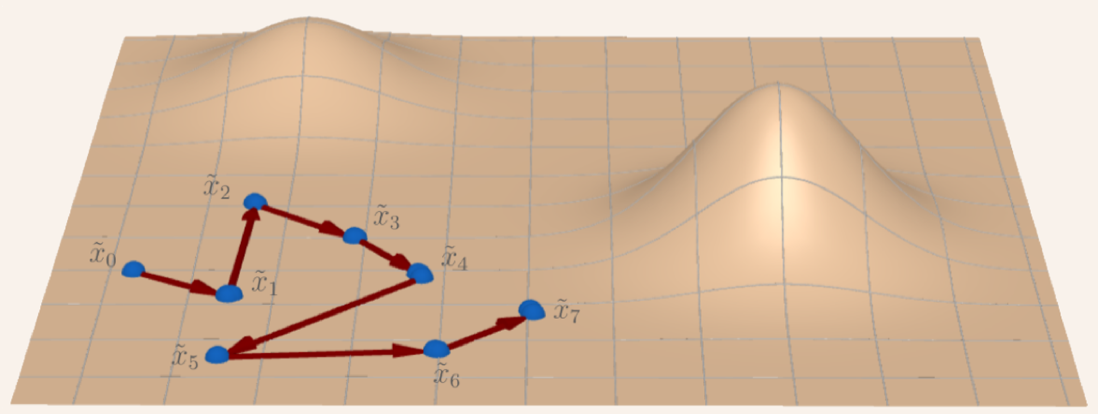

Figure 1: Slow mixing. Even with a good score estimate, Langevin can linger in low-density regions.

Noise-Conditioned Score Networks. Song & Ermon address these obstacles by learning scores across different noisy versions of the data distribution. For each noise scale σ, they consider the distribution q_σ obtained by corrupting data with $\mathcal{N}(0, \sigma^2 I)$. This smoothing spreads mass in the ambient space, making the score well-defined and significantly easier to estimate.

They then train a noise-conditional score network $s_\theta(x, \sigma) \approx \nabla_x \log q_\sigma(x)$ using denoising score matching at each noise scale.

At generation time, sampling follows a coarse-to-fine *annealing* schedule over σ: large noise levels first enable fast global exploration and mode switching, while progressively smaller σ values sharpen the sample and guide it back toward the data manifold.

Stochastic Interpolants

Stochastic interpolants provide a unifying framework connecting deterministic dynamics (flow matching) and stochastic dynamics (diffusion). They describe a random path between two samples x_0 and x_1 as the sum of a deterministic interpolant and a Gaussian term modulated by a function $\gamma(t)$. This formalism explicitly separates geometric transport from randomness and allows precise control over the trade-off between fidelity, diversity, and numerical stability via the diffusion coefficient ϵ.

Given two probability density functions $\rho_0, \rho_1 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, a *stochastic interpolant* between ρ_0 and ρ_1 is a stochastic process $(x_t)_{t\in[0,1]}$ defined by

$$x_t = I(t, x_0, x_1) + \gamma(t)\, z,$$

where $I \in C^2([0,1], C^2(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^d))$ satisfies the boundary conditions $(I(0, x_0, x_1) = x_0,\ I(1, x_0, x_1) = x_1)$ and , $\gamma \in C^2([0,1], \mathbb{R}^+)$ satisfies $\gamma(0) = \gamma(1) = 0$.
The key points here is that by defining :

$$b(t,x) = \mathbb{E}[\dot{x}_t \mid x_t = x] = \mathbb{E}[\partial_t I(t, x_0, x_1) + \dot{\gamma}(t)\, z \,|\, x_t = x]$$
$$s(t,x) = \nabla \log \rho(t,x) = -\,\gamma(t)^{-1}\, \mathbb{E}[z \,|\, x_t = x]$$

then if we denote $\rho(t,x)$ the law of x_t, one can show that for any $\varepsilon \in C^0([0,1])$ with $\varepsilon(t) \geq 0$ for all $t \in [0,1]$.

1. **Forward Fokker–Planck equation**

$$\begin{cases} \partial_t \rho + \nabla \cdot (b_F \rho) = \varepsilon(t)\, \Delta\rho, \\ \rho(0,\cdot) = \rho_0, \end{cases}$$

where the forward drift is defined by

$$b_F(t,x) = b(t,x) + \varepsilon(t)\, s(t,x).$$

2. **Backward Fokker–Planck equation**

$$\begin{cases} \partial_t \rho + \nabla \cdot (b_B \rho) = -\,\varepsilon(t)\, \Delta\rho, \\ \rho(1,\cdot) = \rho_1, \end{cases}$$

where the backward drift is defined by

$$b_B(t,x) = b(t,x) - \varepsilon(t)\, s(t,x).$$

It is important to note that b and s can be compute by minimizing an expectation over parameters x_0, x_1, z_1. Then by training a model over this least square error, we can use the output as a generative modele :

Stochastic Interpolants (2/2)

At any $t \in [0,1]$, the law of the stochastic interpolant x_t coincides with:

- **Probability flow:** $\frac{d}{dt} X_t = b(t, X_t)$, forward from $X_0 \sim \rho_0$ or backward from $X_1 \sim \rho_1$.
- **Forward SDE:** $dX_t^F = b_F(t, X_t^F)\, dt + 2\,\varepsilon(t)\, dW_t$, $X_0^F \sim \rho_0$.
- **Backward SDE:** $dX_t^B = b_B(t, X_t^B)\, dt + 2\,\varepsilon(t)\, dW_t^B$, $X_1^B \sim \rho_1$, $W_t^B = -W_{1-t}$,

Then using the same train b and s we vary ϵ to generate sample from ρ_0 which at the end follow ρ_1. A non-zero $\epsilon(t)$ adds stochasticity, increasing sample diversity, stabilizing training, and linking to Schrödinger bridges. However, it may blur details and reduce trajectory fidelity, requiring careful tuning of the noise level.

Experiments

Setup. CelebA 64×64, UNet $\approx$10M params. We test interpolants $I(t)$ (linear / trig / enc–dec), noise schedules $\gamma(t)$, and diffusion levels $\epsilon \in \{0, 0.25, 0.5\}$ (RK4, 50 steps).

Qualitative findings (mask-to-image).

- **Large diffusion hurts fidelity:** $\epsilon = 0.5$ increases stochasticity but often induces *global drift* and *over-smoothing* (limited model capacity).
- **Best trade-off at moderate diffusion:** $\epsilon = 0.25$ is more stable and preserves identity better, but can still *lose fine details.*
- **Dataset prior shows up with $\epsilon > 0$:** stochastic trajectories favor more *plausible* (non-uniform) backgrounds, sometimes reducing strict mask fidelity.
- **Time scheduling matters:** trig schedules refine endpoints (better boundary accuracy) but may *under-explore* mid-time changes.