Score-Based Generative Models.

A Historical and Conceptual Route Toward Convergence Results

Stanislas Strasman

June 2, 2026

Sorbonne University, Laboratoire de Probabilités, Statistique et Modélisation (LPSM)

Generative AI

Data

We observe a **finite** dataset

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \pi_{\text{data}}\,, \qquad \pi_{\text{data}} \in \mathcal{P}(\mathbb{R}^d) \quad \text{unknown}\,.$$

Unsupervised Machine Learning

Learn a sampling mechanism whose output law $\widehat{\pi}$ satisfies

$$\widehat{\pi} \approx \pi_{\text{data}}\,.$$

⚠ both a learning and a sampling problem.

What does $\approx$ mean?

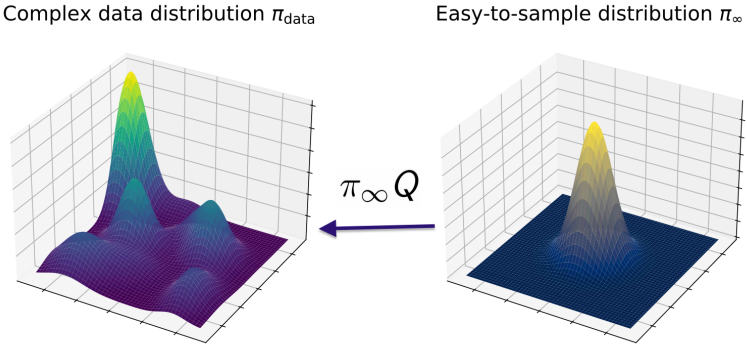
The notion of approximation depends on the application: visual quality, downstream performance, or proper mathematical distance on the space of probability measures.

Generative Modeling as a Mathematical Problem

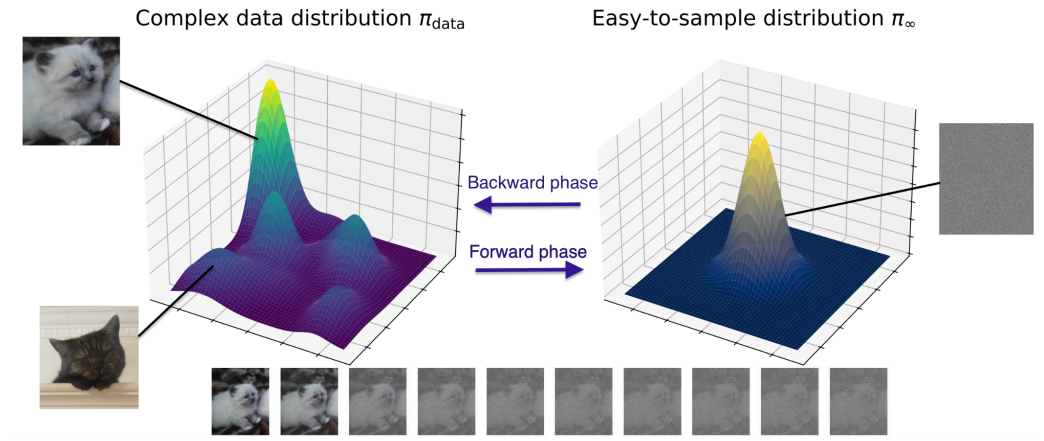
Simuable generation model

Choose an easy-to-sample prior π_∞ and a simulable transition kernel Q such that

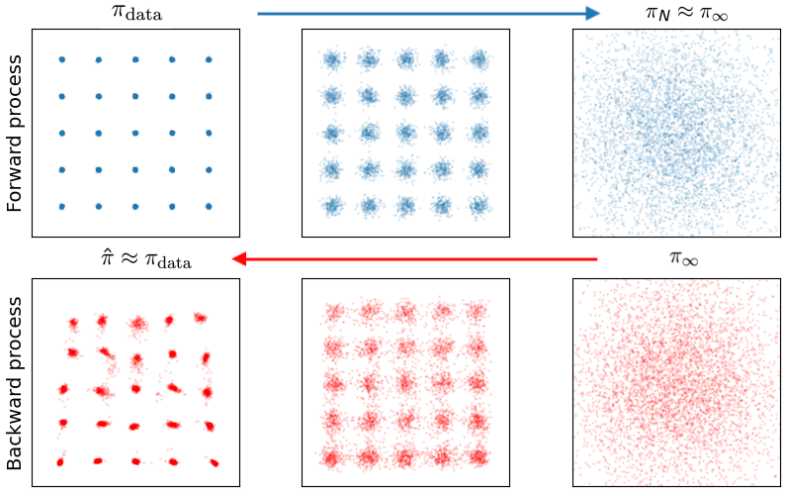
$$\widehat{\pi} = \pi_\infty Q \qquad \text{and} \qquad \pi_\infty Q \approx \pi_{\text{data}}.$$



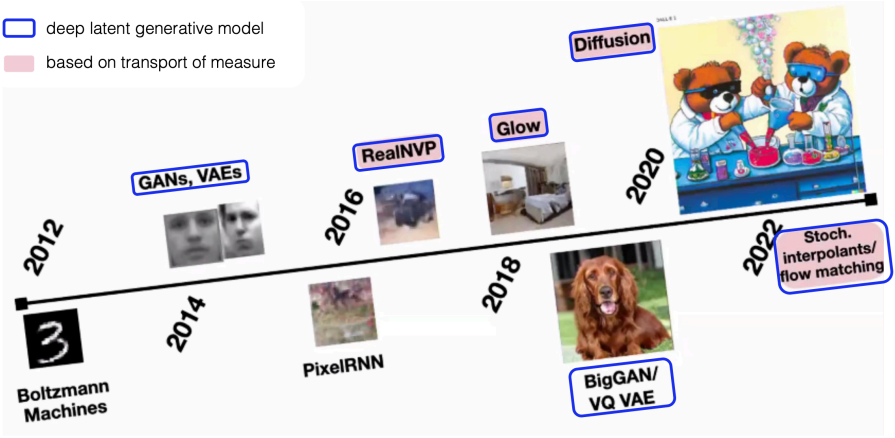
Diffusion model philosophy



Synthetic example: 2-dimensional mixture of 25 Gaussian.

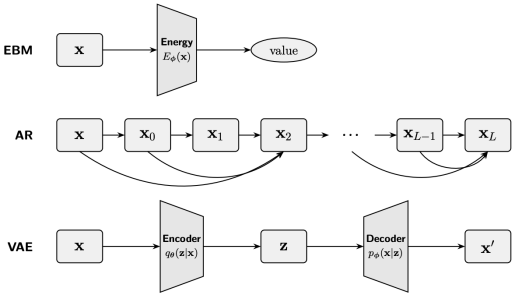


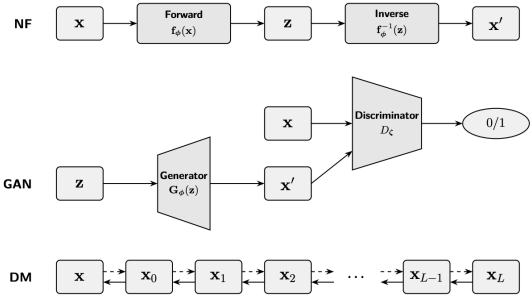
Historical Development of Generative AI



Slide credit: Michael Albergo

Classical Cartography of Generative Models





Source: The Principles of Diffusion Models (2025)

Roadmap

Section 1: a brief history of score-based generative models.

a) A sampling mechanism: the Langevin equation.

b) A learning mechanism: score matching.

c) Gaussian smoothing, denoising, and annealing.

d) Time reversal: a measure transport perspective.

Section 2: SGMs in practice, convergence and stability.

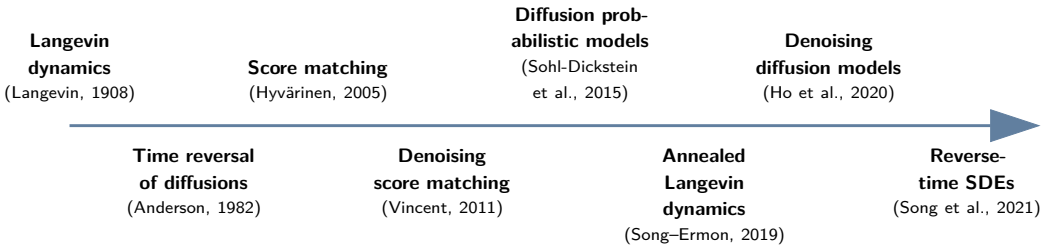
a) The three approximations in SGMs.

b) Error analysis.

Selected research projects

Section 1: a brief history of score-based generative models.

A conceptual route: Langevin sampling, score matching, and time reversal.



Not only an ML story

From this viewpoint, diffusion models are rooted in stochastic sampling, score matching, and time reversal of Markov diffusion processes.

Both their learning procedure and their sampling mechanism revolve around one central quantity: the score function.

The Langevin Equation (1908)

Suppose the target distribution π_{data} admits a density writen as:

$$p_0(x) := \frac{1}{Z} \exp\{-U(x)\}\,, \qquad Z := \int_{\mathbb{R}^d} \exp\big(-U(x)\big)\mathrm{d}x \in (0,\infty)\,.$$

Langevin diffusion

For some $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$

$$\mathrm{d}X_t = \nabla \log p_0(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \qquad X_0 \sim \mu_0\,. \tag{1}$$

- $(X_t)_{t\geq 0}$ defines a Markov diffusion process.
- Under suitable assumptions on U, π_{data} is invariant and

$$\mathcal{L}(X_t) \to \pi_{\text{data}} \qquad \text{as } t \to \infty\,.$$

Informal Interpretation: Noisy Gradient Ascent

If we remove the noise, (1) becomes

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = \nabla \log p_0(X_t)\,. \tag{2}$$

Hence, the deterministic part is

$$\text{gradient ascent on } \log p_0 \quad \Longleftrightarrow \quad \text{gradient descent on } U\,.$$

Pure **gradient ascent** is not a sampler: it collapses to one local mode. Noise adds **random exploration**.

$$\sqrt{2}\int_0^t \mathrm{d}B_s \sim \mathcal{N}(0, 2t\mathrm{I}_d)\,. \tag{3}$$

can be viewed as the limit of a discrete time process

$$X_{t_{k+1}} - X_{t_k} = \sqrt{2h}\, Z_{k+1}\,, \qquad Z_{k+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{I}_d)\,, \quad h \to 0\,.$$

- (2) moves particles toward high-density regions.
- (3) prevents collapse and helps explore the distribution.

From Langevin to ULA

Unadjusted Langevin Algorithm

Euler–Maruyama discretization with step size $h > 0$:

$$X_{k+1} = X_k + h\nabla \log p_0(X_k) + \sqrt{2h}\,\xi_{k+1}, \qquad \xi_{k+1} \sim \mathcal{N}(0, \mathrm{I}_d).$$

Simple example: Ornstein-Uhlenbeck Process. For, $p_0 = \mathcal{N}(0, \mathrm{I}_d)$

$$U(x) = \frac{\|x\|^2}{2},$$

ULA gives

$$X_{k+1} = (1-h)X_k + \sqrt{2h}\,\xi_{k+1}.$$

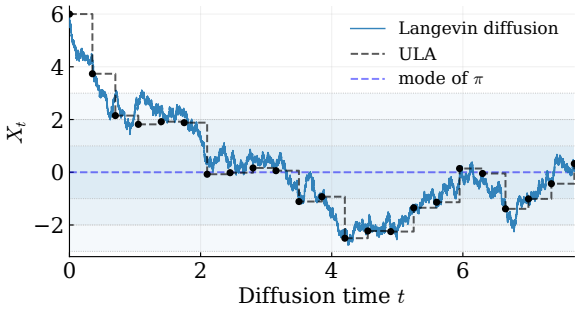
Starting from $X_0 = x$,

$$X_k = (1-h)^k x + \sqrt{2h}\sum_{j=1}^{k}(1-h)^{k-j}\xi_j \quad \Rightarrow \quad X_k \sim \mathcal{N}\left((1-h)^k x, \frac{2}{2-h}\big(1-(1-h)^{2k}\big)\mathrm{I}_d\right).$$

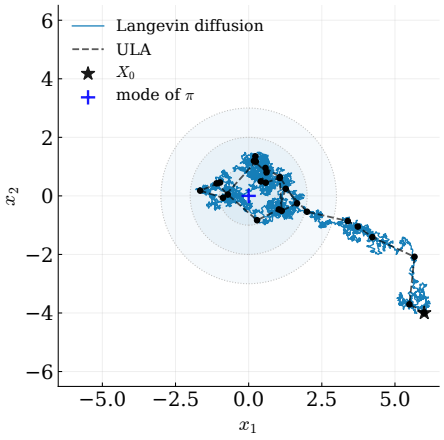
If $0 < h < 2$, then

$$X_k \Rightarrow \pi_h^{\mathrm{ULA}} = \mathcal{N}\left(0, \frac{2}{2-h}\mathrm{I}_d\right) \quad \text{as } k \to \infty.$$

Langevin Diffusion and ULA: Gaussian setting



1-d example targeting $\mathcal{N}(0, 1)$



2-d example targeting $\mathcal{N}(0, \mathrm{I}_2)$

Sampling is not yet Generative Modeling

Langevin sampling assumes

$$\nabla \log p_0(x) \quad \text{is known.}$$

Then one can simulate

$$X_{k+1} = X_k + h\nabla \log p_0(X_k) + \sqrt{2h}\xi_{k+1}.$$

Generative modeling gives

$$X_1, \dots, X_n \sim \pi_{\text{data}},$$

but no explicit formula for

$$p_0 \qquad \text{or} \qquad \nabla \log p_0.$$

A natural idea is to learn a parametric model $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$ by minimizing the Fisher divergence

$$\mathcal{L}_{\text{explicit}}(\theta) = \frac{1}{2}\mathbb{E}_{X\sim\pi_{\text{data}}}\left[\|s_\theta(X) - \nabla \log p_0(X)\|^2\right].$$

⚠ $\nabla \log p_0(X)$ remains unknown!

Hyvärinen's Score Matching (2005)

We would like to fit the unknown score through

$$\mathcal{L}_{\text{explicit}}(\theta) = \frac{1}{2}\mathbb{E}_{X\sim p_0}\left[\|\mathsf{s}_\theta(X) - \nabla \log p_0(X)\|^2\right] .$$

Hyvärinen's trick

Expand the square:

$$\mathcal{L}_{\text{explicit}}(\theta) = \frac{1}{2}\int p_0\|\mathsf{s}_\theta\|^2 \mathrm{d}x - \int \langle \mathsf{s}_\theta, \nabla p_0\rangle \mathrm{d}x + C.$$

Under integration-by-parts conditions,

$$\int_{\mathbb{R}^d} \langle \mathsf{s}_\theta(x), \nabla p_0(x)\rangle \mathrm{d}x = -\int_{\mathbb{R}^d} \mathrm{div}(\mathsf{s}_\theta)(x)\, p_0(x)\mathrm{d}x.$$

Hence, up to a constant independent of θ,

$$\boxed{\mathcal{L}_{\text{SM}}(\theta) = \mathbb{E}_{X\sim p_0}\left[\frac{1}{2}\|\mathsf{s}_\theta(X)\|^2 + \mathrm{div}(\mathsf{s}_\theta)(X)\right].} \tag{4}$$

Direct Score Matching: Trainable but Expensive

Monte Carlo approximation

Given data samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \pi_{\text{data}}$, (4) can be approximated by

$$\widehat{\mathcal{L}}_{\text{SM},n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{2} \|\mathsf{s}_\theta(X_i)\|^2 + \operatorname{div} \mathsf{s}_\theta(X_i) \right].$$

High-dimensional bottleneck

$$\operatorname{div} \mathsf{s}_\theta(x) = \operatorname{Tr}\left(\nabla_x \mathsf{s}_\theta(x)\right)$$

Computing $\operatorname{div} \mathsf{s}_\theta(x)$ is expensive for large neural networks and high-dimensional data.

For a 64×64 RGB image,

$$d = 64 \cdot 64 \cdot 3 = 12288.$$

$\operatorname{div} \mathsf{s}_\theta(x)$ involves 12288 input-derivative terms per image, and then differentiating again with respect to θ.

The Conditioning Trick I: Key Identity

Introduce a known corruption kernel

Draw a data point and corrupt it through a known kernel:

$$X_0 \sim \pi_{\mathrm{data}}, \qquad \tilde{X} \mid X_0 = x_0 \sim \tilde{q}(\cdot \mid x_0),$$

and suppose that the conditional score is known

$$\boxed{Y := \nabla_{\tilde{x}} \log \tilde{q}(\tilde{X} \mid X_0)}$$

Conditional score identity

Then, $\tilde{X}$ has marginal density $\tilde{p}(x) = \int_{\mathbb{R}^d} \tilde{q}(x \mid x_0)\, \pi_{\mathrm{data}}(\mathrm{d}x_0)$. Since $\tilde{q}$ is known, under regularity assumptions,

$$\nabla \tilde{p}(x) = \int_{\mathbb{R}^d} \tilde{q}(x \mid x_0) \nabla_x \log \tilde{q}(x \mid x_0)\, \pi_{\mathrm{data}}(\mathrm{d}x_0).$$

Therefore,

$$\nabla \log \tilde{p}(\tilde{X}) = \mathbb{E}\big[\nabla_{\tilde{x}} \log \tilde{q}(\tilde{X} \mid X_0) \,\big|\, \tilde{X}\big] = \mathbb{E}\left[Y \middle| \tilde{X}\right].$$

The Conditioning Trick II: Regression View

$$\boxed{Y := \nabla_{\tilde{x}} \log \tilde{q}(\tilde{X} \mid X_0)}$$

Conditional expectation as an L^2 projection

The best prediction of Y from $\tilde{X}$ is $f^\star(\tilde{X}) = \mathbb{E}[Y \mid \tilde{X}]$. Equivalently,

$$f^\star \in \operatorname*{arg\,min}_{f:\mathbb{R}^d \to \mathbb{R}^d} \mathbb{E}\left[\|f(\tilde{X}) - Y\|^2\right].$$

By the previous identity, $f^\star(\tilde{X}) = \nabla \log \tilde{p}(\tilde{X})$.

Conditional score matching

We train s_θ by minimizing the fully computable loss

$$\mathcal{L}_{\mathrm{CSM}}(\theta) = \frac{1}{2}\mathbb{E}\left[\left\|s_\theta(\tilde{X}) - \nabla_{\tilde{x}} \log \tilde{q}(\tilde{X} \mid X_0)\right\|^2\right].$$

⚠ learns the score of $\tilde{X}$ not of X_0 !

Vincent's Denoising Score matching I (2011)

For $\sigma > 0$, define

$$\tilde{X}_\sigma = X_0 + \sigma Z, \qquad X_0 \sim \pi_{\text{data}}, \qquad Z \sim \mathcal{N}(0, \mathrm{I}_d), \qquad Z \perp\!\!\!\perp X_0.$$

Equivalently,

$$q_\sigma(x \mid x_0) = (2\pi\sigma^2)^{-d/2} \exp\left(-\frac{\|x - x_0\|^2}{2\sigma^2}\right).$$

Closed-form conditional score

The conditional score is explicit:

$$\nabla_x \log q_\sigma(x \mid x_0) = -\frac{x - x_0}{\sigma^2}.$$

Fast sampling

Gaussian sampling highly optimized in modern library through Ziggurat methods (Numpy) or Box–Muller transforms (Pytorch).

Denoising Score matching II

Regularization by smoothing

The corrupted marginal has density

$$p_\sigma(x) = \int_{\mathbb{R}^d} \varphi_\sigma(x - x_0)\, \pi_{\mathrm{data}}(\mathrm{d}x_0), \qquad \varphi_\sigma = \mathcal{N}(0, \sigma^2 \mathrm{I}_d).$$

Even if π_{data} is singular p_σ is smooth for every $\sigma > 0$.

Denoising interpretation

Using the Gaussian conditional score,

$$\mathcal{L}_{\mathrm{DSM}}(\theta; \sigma) = \frac{1}{2}\mathbb{E}\left[\left\| \mathsf{s}_\theta(\tilde{X}_\sigma, \sigma) + \frac{\tilde{X}_\sigma - X_0}{\sigma^2} \right\|^2\right].$$

Equivalently, with

$$D_\theta(y, \sigma) := y + \sigma^2 \mathsf{s}_\theta(y, \sigma),$$

one obtains

$$\mathcal{L}_{\mathrm{DSM}}(\theta; \sigma) = \frac{1}{2\sigma^4}\mathbb{E}\left[\|D_\theta(\tilde{X}_\sigma, \sigma) - X_0\|^2\right].$$

U-Net Architecture for Denoising

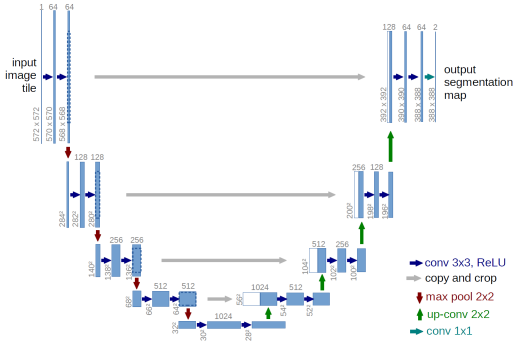


Image credit: S. Mallat, lecture notes, 2024.

Takeaway

The success of diffusion models is tied to efficient denoising architectures.

The fixed-noise trade-off

Recall that

$$\tilde{X}_\sigma = X_0 + \sigma Z, \qquad Z \sim \mathcal{N}(0, I_d), \qquad \pi_\sigma := \mathcal{L}(\tilde{X}_\sigma)$$

How to select σ ?

Small σ: low bias

- The corrupted law remains close to the data.

$$\begin{aligned} \mathcal{W}_2^2(\pi_{\text{data}}, \pi_\sigma) &\leq \mathbb{E}\|X_0 - \tilde{X}_\sigma\|^2 \\ &= \sigma^2 \mathbb{E}\|Z\|^2 \\ &= \sigma^2 d \end{aligned}$$

Too small σ: hard target and mixing

- The conditional score maybe poorly scaled

$$\begin{aligned} &\mathbb{E}\left[\left\|\nabla_x \log q_\sigma(\tilde{X}_\sigma \mid X_0)\right\|^2\right] \\ &= \mathbb{E}\left\|\frac{Z}{\sigma}\right\|^2 = \frac{d}{\sigma^2}. \end{aligned}$$

- Sharp barriers may remain, difficult mixing and learning.

One fixed noise level is not enough.

A Simple Two-Cluster Toy Example

Data distribution

In one dimension, consider

$$\pi_{\text{data}} = \frac{1}{2}\delta_{-a} + \frac{1}{2}\delta_{+a}, \qquad a > 0.$$

After Gaussian smoothing with variance σ^2,

$$\tilde{X}_\sigma = X_0 + \sigma Z, \qquad Z \sim \mathcal{N}(0,1),$$

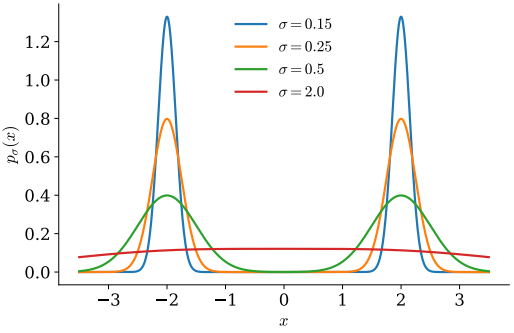
the smoothed density becomes remains multimodal if $\sigma \ll a$, with φ the Gaussian density

$$p_\sigma(x) = \frac{1}{2}\varphi_\sigma(x-a) + \frac{1}{2}\varphi_\sigma(x+a).$$

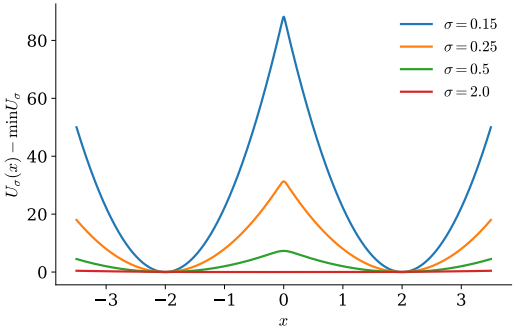
Takeaway

Small Gaussian smoothing does not remove multimodality: the two modes remain separated by a sharp barrier, so Langevin-type sampling may still mix poorly.

Gaussian Smoothing of a Two-Cluster Distribution

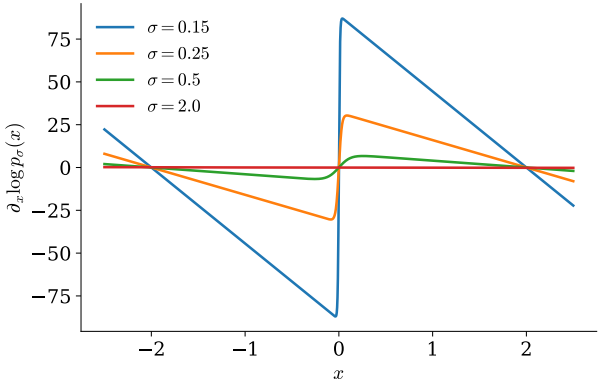


Density p_σ



Potential $U_\sigma = -\log p_\sigma$

For $a = 2$, small noise preserves multimodality and produces sharp barriers.

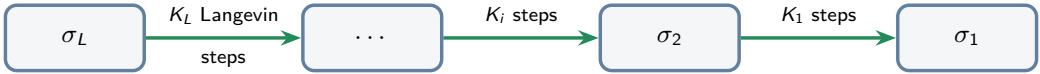


Score function $\partial_x \log p_\sigma$

Multi-scale learning and Annealed Langevin Dynamics (Song, 2019)

Multi-scale idea

Choose a decreasing noise ladder $\sigma_L > \sigma_{L-1} > \cdots > \sigma_1$ and learn scores $\mathsf{s}_\theta(\cdot, \sigma_i) \approx \nabla \log p_{\sigma_i}$.



Frozen Langevin steps at each level

For $i = L, L-1, \ldots, 1$, choose step sizes $\eta_i > 0$ and run K_i Langevin steps

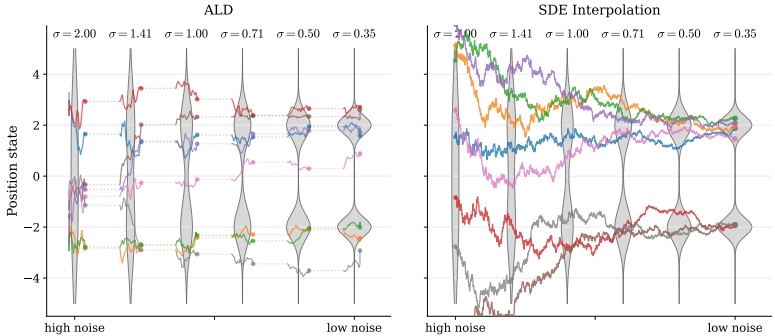
$$x_{i,k+1} = x_{i,k} + \eta_i \mathsf{s}_\theta(x_{i,k}, \sigma_i) + \sqrt{2\eta_i}\, \xi_{i,k+1}, \qquad \xi_{i,k+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{I}_d).$$

Then use the output at level σ_i to initialize the next level.

Main limitations

- Sequential sampler: errors are passed down the ladder.
- Computationnally expensive $\sum_{i=1}^{L} K_i$ steps.

ALD vs transport of measure



Takeaway

The score is no longer used only for local equilibration; it becomes part of a reverse-time generative process.

From Marginals to Path Space

What we learned so far

Score matching gives access to scores $\nabla \log p_\sigma$ of Gaussian-corrupted variables:

$$X_\sigma = X_0 + \sigma Z, \qquad Z \sim \mathcal{N}(0, \mathrm{I}_d), \qquad X_0 \sim \pi_{\mathrm{data}}\,.$$

New viewpoint

Instead of determining a sequence of marginals, choose a stochastic process whose marginals realize them.

VE path-space realization

The Brownian noising process

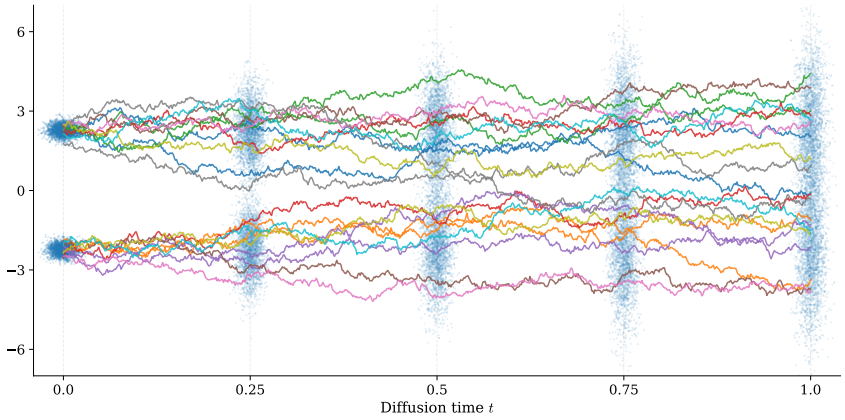
$$\mathrm{d}\overrightarrow{X}_t = \sqrt{2}\,\mathrm{d}B_t, \qquad \overrightarrow{X}_0 \sim \pi_{\mathrm{data}},$$

satisfies

$$\overrightarrow{X}_t = X_0 + \sqrt{2t}\,Z, \qquad \overrightarrow{X}_t \mid X_0 = x_0 \sim \mathcal{N}(x_0, 2t\,\mathrm{I}_d).$$

Thus it realizes Gaussian smoothing with $\sigma_t^2 = 2t$.

Gaussian Smoothing as a Path-Space Process



Blue clouds: Gaussian marginals $\overrightarrow{X}_t \mid \overrightarrow{X}_0 = x_0 \sim \mathcal{N}(x_0, 2t)$.
Orange curves: sample paths of a VE process $\mathrm{d}\overrightarrow{X}_t = \sqrt{2}\,\mathrm{d}B_t$.

Time Reversal of diffusion processes (Anderson (1982))

Let p_t be the time-marginal density of $(\overrightarrow{X}_t)_{t\in[0,T]}$ solution to $\mathrm{d}\overrightarrow{X}_t = \sqrt{2}\,\mathrm{d}B_t, \quad \overrightarrow{X}_0 \sim \pi_{\mathrm{data}}.$

Reverse-time process

The time-reversed process for $T > 0$ fixed

$$(\overleftarrow{X}_t)_{t\in[0,T]} \overset{\mathcal{L}}{=} (\overrightarrow{X}_{T-t})_{t\in[0,T]}$$

is again a diffusion process and satisfies

$$\mathrm{d}\overleftarrow{X}_t = 2\nabla \log p_{T-t}(\overleftarrow{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad \overleftarrow{X}_0 \sim p_T. \tag{5}$$

Generative AI via transport of measure

If the score $\nabla \log p_t$ is known or learned, then generation amounts to:

$$\text{sample } \overleftarrow{X}_0 \sim p_T \quad \longrightarrow \quad \text{simulate the reverse SDE (5)} \quad \longrightarrow \quad \overleftarrow{X}_T \sim \pi_{\mathrm{data}}.$$

Time reversal is not frozen Langevin

Forward Brownian noising

$$\mathrm{d}\overrightarrow{X}_t = \sqrt{2}\,\mathrm{d}B_t, \qquad \overrightarrow{X}_0 \sim \pi_{\mathrm{data}}.$$

True reverse dynamics

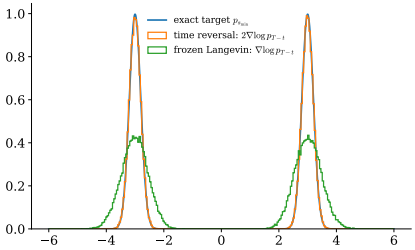
$$\mathrm{d}\overleftarrow{X}_t = 2\nabla \log p_{T-t}(\overleftarrow{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t.$$

transports the marginals backward along the heat flow.

Time inhomogeneous Langevin

$$\mathrm{d}Y_t = \nabla \log p_{T-t}(Y_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t.$$

makes p_t invariant when time is frozen.



Toy experiment

- Blue line: target. Both dynamics start from p_T and use the exact score.
- Orange samples: target the right distribution.
- Green samples: if we would froze t and kept running Langevin, it would converge, but with changing schedule it remains too diffuse.

Section 2: SGMs in practice, convergence and stability.

From the Ideal Reverse Process to a Practical Sampler

Exact reverse process

By time reversal, the ideal reverse dynamics are

$$\mathrm{d}\overleftarrow{X}_t = 2\nabla \log p_{T-t}(\overleftarrow{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad \overleftarrow{X}_0 \sim p_T. \tag{6}$$

If $Q_T = \mathbb{P}(\overrightarrow{X}_T \in \mathrm{d}y | \overrightarrow{X}_0 = x)$, then

$$\boxed{p_T Q_T = \pi_{\mathrm{data}}}$$

Practical obstacles

Initialization

p_T still depends on the data:

$p_T(x) = \mathcal{L}(\overrightarrow{X}_0 + \sqrt{2T}Z)$.

Approximation

The score is not available and must be learned:

$\nabla \log p_{T-t}(x) \approx s_\theta(T-t, x)$.

Discretization

The transisions probabilities of (6) are unknown the dynamics needs to be discretied.

Error I: Initialization Error

Exact versus practical initialization

The ideal reverse process starts from

$$p_T(x) = \int \mathcal{N}(x; y, 2TI_d) p_0(y) \mathrm{d}y$$

but p_T depends on π_{data}. In practice, we use

$$\pi_\infty = \mathcal{N}(0, 2TI_d).$$

Initialization error: $\pi_{\mathsf{data}} \simeq \pi_\infty Q_T$

Short KL argument

By convexity of KL,

$$\mathrm{KL}(p_T \,\|\, \pi_\infty) \leq \int \mathrm{KL}(\mathcal{N}(y, 2TI_d) \,\|\, \mathcal{N}(0, 2TI_d))\, p_0(y)\, \mathrm{d}y = \frac{\mathbb{E}\|\overrightarrow{X}_0\|^2}{4T}.$$

For T large $\overrightarrow{X}_0$ signal gets smaller.

VE noising: 2-d example

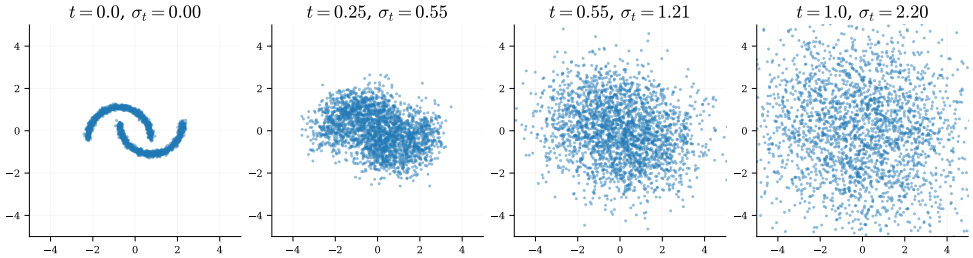


Figure 1: VE: $\overrightarrow{X}_t \overset{\mathcal{L}}{=} \overrightarrow{X}_0 + \sigma_t Z$.

⚠ At $t = 1$, if σ_1 is large compared with the data scale, the marginal law is visually almost indistinguishable from high-variance Gaussian draws.

Error II: Score Approximation

Practical reverse dynamics

$\nabla \log p_{T-t}(\overleftarrow{X}_t)$ is unknown. We learn a neural approximation $s_\theta(x, t)$ and the reverse SDE becomes

$$\mathrm{d}X_t^\theta = 2s_\theta(X_t^\theta, T-t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad X_0^\theta \sim \pi_\infty.$$

Approximation error: $\pi_{\mathrm{data}} \approx \pi_\infty Q_T^\theta$

Denoising score matching

For the Brownian noising process, $\overrightarrow{X}_t = \overrightarrow{X}_0 + \sqrt{2t}Z$ and

$$\nabla_x \log p_{t|0}(\overrightarrow{X}_t \mid \overrightarrow{X}_0) = -\frac{\overrightarrow{X}_t - \overrightarrow{X}_0}{2t}.$$

Thus we train s_θ by minimizing

$$\mathcal{L}_{\mathrm{DSM}}(\theta) = \int_0^T \mathbb{E}\left[\left\| s_\theta(\overrightarrow{X}_t, t) + \frac{\overrightarrow{X}_t - \overrightarrow{X}_0}{2t} \right\|^2\right] \mathrm{d}t = T\mathbb{E}\left[\left\| s_\theta(\overrightarrow{X}_\tau, \tau) + \frac{\overrightarrow{X}_\tau - \overrightarrow{X}_0}{2\tau} \right\|^2\right],$$

for $\tau \sim \mathcal{U}([0, T])$ it can be approximated by Monte Carlo.

Error III: Discretization Error

Continuous approximate reverse SDE

After replacing the true score by s_θ, the reverse dynamics is still continuous-time:

$$\mathrm{d}X_t^\theta = 2s_\theta(X_t^\theta, T-t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad X_0^\theta \sim \pi_\infty.$$

Let Q_T^θ denote its Markov kernel.

Euler–Maruyama sampler

Choose N steps, $h = T/N$, and $t_k = kh$. The practical sampler is

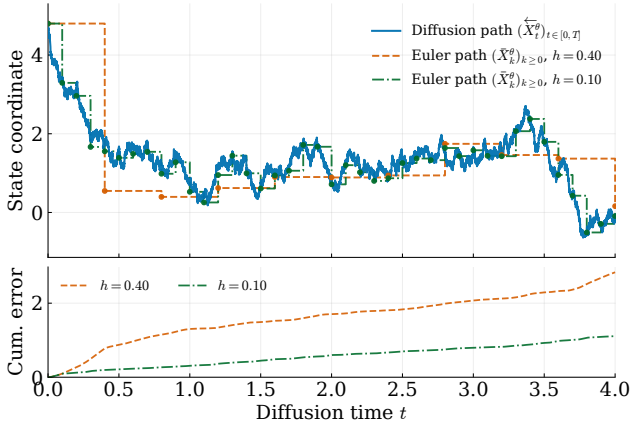
$$X_{k+1}^{\theta,N} = X_k^{\theta,N} + 2h\,s_\theta(X_k^{\theta,N}, T-t_k) + \sqrt{2h}\,Z_{k+1}, \qquad Z_{k+1} \sim \mathcal{N}(0, I_d).$$

Let $\bar{Q}_{T,N}^\theta$ denote the corresponding N-step Markov kernel.

Discretization error:

$$\pi_{\mathrm{data}} \approx \pi_\infty \bar{Q}_{T,N}^\theta := \hat{\pi}_{\infty,N}^\theta$$

1-d pathwise intuition



All discretizations are coupled with the same Brownian path and $h = T/N$.

Score-Based Generative Modeling: Algorithm

Training

For $i = 1, \dots, B$, sample

$$X_0^i \sim \pi_{\text{data}}, \qquad \tau_i \sim \mathcal{U}((0, T]), \qquad Z_i \sim \mathcal{N}(0, \mathrm{I}_d),$$

and set

$$X_{\tau_i}^i = X_0^i + \sqrt{2\tau_i}\, Z_i.$$

The conditional-score target is

$$r_i := -\frac{X_{\tau_i}^i - X_0^i}{2\tau_i}.$$

Minimize the minibatch loss

$$\widehat{\mathcal{L}}_B(\theta) = \frac{1}{2B} \sum_{i=1}^{B} \left\| \mathrm{s}_\theta(X_{\tau_i}^i, \tau_i) - r_i \right\|^2.$$

SGD update:

$$\theta_{m+1} = \theta_m - \gamma_m \nabla_\theta \widehat{\mathcal{L}}_B(\theta_m).$$

Sampling

Initialize from the high-noise prior

$$Y_0 \sim \pi_\infty = \mathcal{N}(0, 2T\mathrm{I}_d).$$

Let

$$h = \frac{T}{N}, \qquad t_k = kh.$$

For $k = 0, \dots, N-1$, simulate

$$Y_{k+1} = Y_k + 2h\, \mathrm{s}_{\hat{\theta}}(Y_k, T - t_k) + \sqrt{2h}\, \xi_{k+1},$$

with

$$\xi_{k+1} \sim \mathcal{N}(0, \mathrm{I}_d).$$

Return

$$Y_N \sim \widehat{\pi}_{\infty, N}^{\hat{\theta}}.$$

The Three SGM Errors

From the ideal law to the practical sampler

The exact reverse process satisfies

$$\pi_{\mathrm{data}} = p_T Q_T.$$

The practical generated law is

$$\widehat{\pi}^{\theta}_{\infty,N} := \pi_\infty \bar{Q}^{\theta}_{T,N}.$$

Error decomposition

For a probability distance d, schematically:

$$\mathrm{d}\left(\pi_{\mathrm{data}}, \widehat{\pi}^{\theta}_{\infty,N}\right) = \mathrm{d}\left(p_T Q_T, \pi_\infty \bar{Q}^{\theta}_{T,N}\right)$$

$$\leq \underbrace{\mathrm{d}(p_T Q_T, \pi_\infty Q_T)}_{\text{initialization}} + \underbrace{\mathrm{d}\left(\pi_\infty Q_T, \pi_\infty Q^{\theta}_T\right)}_{\text{score approximation}} + \underbrace{\mathrm{d}\left(\pi_\infty Q^{\theta}_T, \pi_\infty \bar{Q}^{\theta}_{T,N}\right)}_{\text{discretization}}.$$

Existing Results I: KL and Total Variation

Score-error proxy

We write M for an integrated L^2-score error, schematically

$$M^2 := \int_{\varepsilon}^{T} \mathbb{E}_{\nu_t}\left[\|\mathsf{s}_\theta(X,t) - \nabla \log p_t(X)\|^2\right] \mathrm{d}t < \infty\,,$$

where the law ν_t depends on the precise theorem.

Typical KL bound

For relative entropy,

$$\mathrm{KL}\left(\pi_{\mathrm{data}} \,\middle\|\, \widehat{\pi}^{\theta}_{\infty,N}\right) \lesssim c_1 e^{-T} + c_2 M^2 + c_3 h, \qquad h = \frac{T}{N}.$$

Total variation consequence (Pinsker's inequality)

$$\left\|\pi_{\mathrm{data}} - \widehat{\pi}^{\theta}_{\infty,N}\right\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2}\mathrm{KL}\left(\pi_{\mathrm{data}} \,\middle\|\, \widehat{\pi}^{\theta}_{\infty,N}\right)}.$$

Examples of KL/TV analyses include De Bortoli, Thornton, Heng and Doucet (2021), Conforti, Durmus and Gentiloni Silveri (2023), Chen, Lee and Lu (2023), and Benton, De Bortoli, Doucet and Deligiannidis

Existing Results II: Wasserstein and Stability

The 2-Wasserstein distance

For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \gamma(\mathrm{d}x, \mathrm{d}y).$$

Typical $\mathcal{W}_2$ bound

For OU/VP-type analyses, one often obtains schematically

$$\mathcal{W}_2\left(\pi_{\mathrm{data}}, \widehat{\pi}^{\theta}_{\infty,N}\right) \lesssim c_1 e^{-T} + c_2 M + c_3 \sqrt{h}.$$

Examples of $\mathcal{W}_2$ convergence analyses include Gao, Nguyen and Zhu (2025), Gentiloni-Silveri and Ocello (2025), and related Wasserstein analyses of kinetic and non-log-concave SGMs by Strasman et al. (2025).

Selected references: foundations and algorithms

- Langevin (1908). *Sur la théorie du mouvement brownien*.
- Anderson (1982). *Reverse-time diffusion equation models*.
- Hyvärinen (2005). *Estimation of non-normalized statistical models by score matching*.
- Vincent (2011). *A connection between score matching and denoising autoencoders*.
- Sohl-Dickstein et al. (2015). *Deep unsupervised learning using nonequilibrium thermodynamics*.
- Song and Ermon (2019). *Generative modeling by estimating gradients of the data distribution*.
- Ho, Jain and Abbeel (2020). *Denoising diffusion probabilistic models*.
- Song et al. (2021). *Score-based generative modeling through stochastic differential equations*.

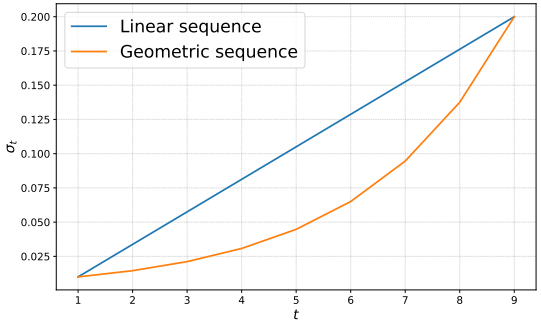
Selected references: convergence and error analysis

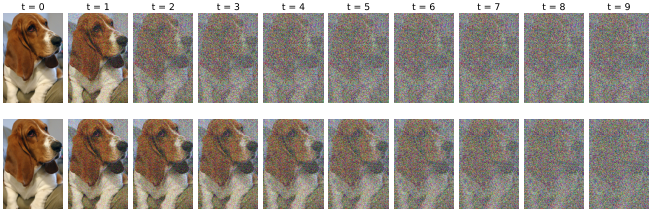
- **KL / TV analyses**
  - De Bortoli, Thornton, Heng and Doucet (2021). *Diffusion Schrödinger bridge with applications to score-based generative modeling*.
  - Conforti, Durmus and Gentiloni Silveri (2025).*KL convergence guarantees for score diffusion models under minimal data assumptions*.
  - Chen, Lee and Lu (2023). *Improved analysis of score-based generative modeling*.
  - Benton, De Bortoli, Doucet and Deligiannidis (2024). *Nearly d-linear convergence bounds for diffusion models via stochastic localization*.

- $\mathcal{W}_2$ **analyses**
  - Gao, Nguyen and Zhu (2025). *Wasserstein convergence guarantees for a general class of score-based generative models*.
  - Gentiloni Silveri and Ocello (2025). *Beyond Log-Concavity and Score Regularity*.
  - Strasman et al. (2025). *An analysis of the noise schedule for score-based generative models*.
  - Strasman, Surendran et al. (2026). *Wasserstein convergence of critically damped langevin diffusions*.

Selected research projects

Study of the noise schedule: can we tell which is better? (Strasman et al., 2025)





Noise schedules as time changes

VE noising with a schedule

A noise schedule is a positive function

$$\beta : [0, T] \to \mathbb{R}_{>0}.$$

It defines a time-changed Brownian noising process

$$\mathrm{d}\overrightarrow{X}_t = \sqrt{2\beta(t)}\,\mathrm{d}B_t, \qquad \overrightarrow{X}_0 \sim \pi_{\mathrm{data}}.$$

Equivalently,

$$\overrightarrow{X}_t = \overrightarrow{X}_0 + B_{2\tau_t}, \qquad \tau_t := \int_0^t \beta(s)\,\mathrm{d}s.$$

Interpretation

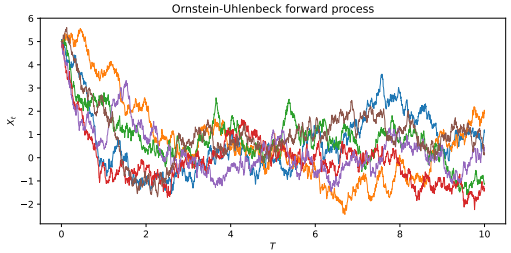
Changing β does not change the basic Gaussian corruption mechanism:

$$\overrightarrow{X}_t \mid \overrightarrow{X}_0 = x_0 \sim \mathcal{N}(x_0, 2\tau_t \mathrm{I}_d).$$

It changes how fast the process moves through the noise levels.

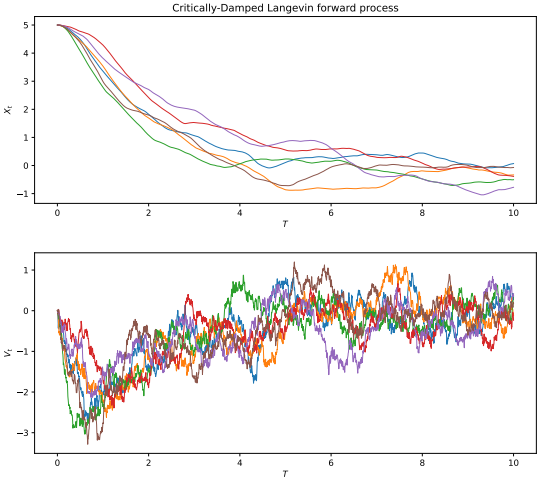
Wassertstein stability of Kinetic SGMs (with Sobihan).

Standard OU noising



Position-only dynamics in $\mathbb{R}^d$.

Critically damped Langevin noising



Phase-space dynamics $(X_t, V_t) \in \mathbb{R}^d \times \mathbb{R}^d$.

Wasserstein Stability of Kinetic SGMs

Critically damped Langevin noising

We enlarge the state space by adding a velocity variable:

$$\overrightarrow{U}_t = \begin{pmatrix} \overrightarrow{X}_t \\ \overrightarrow{V}_t \end{pmatrix} \in \mathbb{R}^d \times \mathbb{R}^d.$$

The forward noising process is

$$\mathrm{d}\begin{pmatrix} \overrightarrow{X}_t \\ \overrightarrow{V}_t \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & \mathrm{I}_d \\ -\mathrm{I}_d & -2\mathrm{I}_d \end{pmatrix}}_{A} \begin{pmatrix} \overrightarrow{X}_t \\ \overrightarrow{V}_t \end{pmatrix} \mathrm{d}t + \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & \sigma \mathrm{I}_d \end{pmatrix}}_{\Sigma} \mathrm{d}B_t, \qquad (\overrightarrow{X}_0, \overrightarrow{V}_0) \sim \pi_{\mathrm{data}} \otimes \pi_v.$$

Main difficulty

Noise is injected only in velocity:

$$\Sigma\Sigma^\top = \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2 \mathrm{I}_d \end{pmatrix}.$$

Stability question

Can we still control the reverse sampler in $\mathcal{W}_2$?

$$\mathcal{W}_2\big(\pi_{\mathrm{data}}, \mathcal{L}(\bar{X}_T^\theta)\big) \lesssim \text{init.} + \text{score} + \text{disc.}$$

The analysis relies on phase-space stability rather than standard elliptic contraction.