SCIENCES SORBONNE UNIVERSITÉ

# Wasserstein-2 Convergence of Critically Damped Diffusion Models

*SGM errors, kinetic extension and stability.*

Stanislas Strasman, S. Surendran, C. Boyer, S. Le Corff, V. Lemaire, A. Ocello.

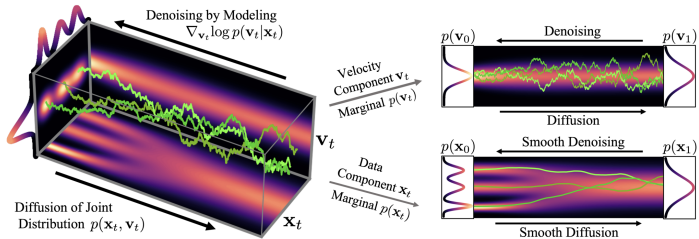
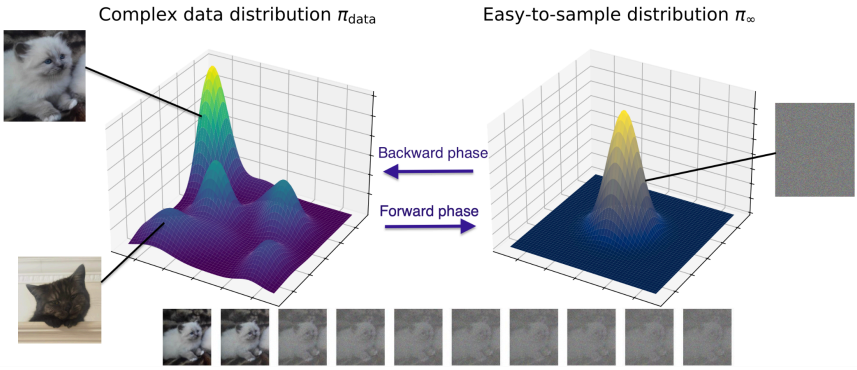



Image credit: Dockhorn (2022)

# Generative Modeling Framework

- $\mathcal{D} = \{X_i\}_{i=1}^n \in (\mathbb{R}^d)^n$ a collection of i.i.d. samples from an unknown distribution $\pi_{\text{data}}$.
- Goal: **generate new samples from** $\pi_{\text{data}}$.



# The Time-Reversal Perspective

Let $p_t$ be the density of the forward process

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

## Time reversal of diffusions

Fix $T > 0$, the backward process

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

is again a diffusion (Anderson, 1982) and solves

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

## Ideal generative identity

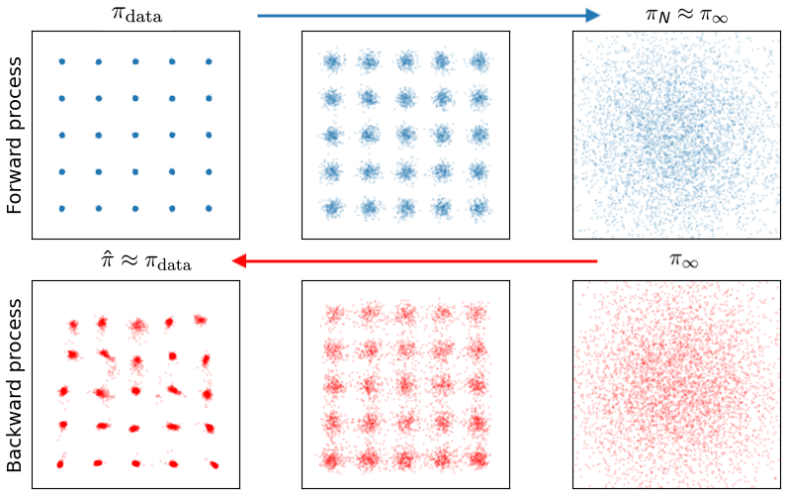
Let $Q_T$ be the Markov transition kernel of the reverse process:

$$Q_T(x, \mathrm{d}y) = \mathbb{P}(\overleftarrow{X}_T \in \mathrm{d}y \mid \overleftarrow{X}_0 = x).$$

Then time reversal gives

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

# Synthetic example: 2-dimensional mixture of 25 Gaussian.



# SGMs in practice: three approximations

⚠ $\boxed{p_T Q_T = \pi_{\text{data}}}$ **is almost generative, but . . .**

## 1. Initialization

$p_T$ depends on the data. The exact terminal law is

$$\overrightarrow{X}_T \overset{\mathcal{L}}{=} \mathrm{e}^{-T}\overrightarrow{X}_0 + \sqrt{1-\mathrm{e}^{-2T}}\, Z$$

For large $T$,

$$p_T \approx \pi_\infty = \mathcal{N}(0, \mathrm{I}_d)$$

$$\pi_{\text{data}} \approx \pi_\infty Q_T$$

## 2. Score approximation

The reverse drift uses the unknown score

$$\nabla \log p_t(x)$$

In practice, learn

$$\mathsf{s}_\theta(x,t) \approx \nabla \log p_t(x)$$

$$\pi_{\text{data}} \approx \pi_\infty Q_T^\theta$$

## 3. Discretization

The reverse SDE should be discretized. Choose

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

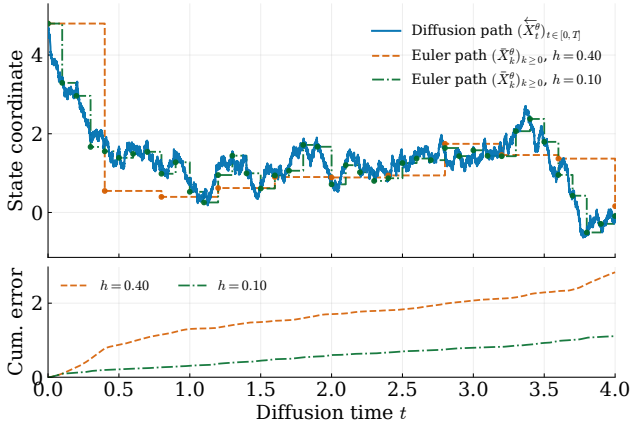
and apply Euler-Maruyama.

$$\pi_{\text{data}} \approx \pi_\infty \bar{Q}^\theta_{T,N}$$

$$\bar{X}^{\theta,N}_{k+1} = \bar{X}^{\theta,N}_k + h\Big(\bar{X}^{\theta,N}_k + 2\mathsf{s}_\theta(\bar{X}^{\theta,N}_k, T-t_k)\Big) + \sqrt{2h}Z_{k+1}\,, \qquad \bar{X}^{\theta,N}_0 \sim \pi_\infty\,, \qquad Z_{k+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$$

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

# 1-d pathwise intuition




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

## The $\mathcal{W}_2$ distance

$$\mathcal{W}_2^2(\pi_{\text{data}}, \widehat{\pi}^{\theta}_{\infty,N}) = \inf\left\{ \mathbb{E}\left[ \left\| \overrightarrow{X}_0 - \bar{X}^{\theta}_{\infty,N} \right\|^2 \right], \ \overrightarrow{X}_0 \sim \pi_{\text{data}}, \bar{X}^{\theta}_{\infty,N} \sim \widehat{\pi}^{\theta}_{\infty,N} \right\}$$

## Score approximation assumption

Assume that,

$$\sup_{0 \le k \le N-1} \left\| \nabla \log p_{T-t_k}\left(\bar{X}_k^{N,\theta}\right) - s_\theta\left(T - t_k, \bar{X}_k^{N,\theta}\right) \right\|_{L^2} \le M\,.$$

## Typical $\mathcal{W}_2$ bound

$$\mathcal{W}_2(\pi_{\text{data}}, \widehat{\pi}^{\theta}_{\infty,N}) \le \underbrace{\mathcal{W}_2(\mathcal{L}(\overleftarrow{X}_T), \mathcal{L}(\bar{X}_N))}_{\textit{Discretization}} + \underbrace{\mathcal{W}_2(\mathcal{L}(\bar{X}_N), \mathcal{L}(\bar{X}_{\infty,N}))}_{\textit{Mixing time}} + \underbrace{\mathcal{W}_2(\mathcal{L}(\bar{X}_{\infty,N}), \mathcal{L}(\bar{X}^{\theta}_{\infty,N})))}_{\textit{Score approx.}}$$

$$\le \mathrm{e}^{-T} c_1 + M c_2 + \sqrt{h} c_3\,,$$

with $T > 0$ the diffusion time, $M$ the score approximation quality and $h = T/N$ the discretization step size.

# Extending the Phase Space: CLD Noising

## Kinetic idea

Augment the data space with a velocity:

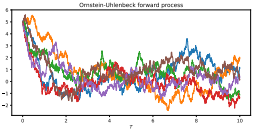
$$\overrightarrow{\mathbf{U}}_t := \begin{pmatrix} \overrightarrow{X}_t \\ \overrightarrow{V}_t \end{pmatrix} \in \mathbb{R}^d \times \mathbb{R}^d, \quad (\overrightarrow{X}_0, \overrightarrow{V}_0) \sim \pi_{\text{data}} \otimes \mathcal{N}(0, \mathrm{I}_d).$$

- Noise is injected only through the velocity.
- $X_t$ and $V_t$ are coupled by the drift.
- Hypoelliptic structure !

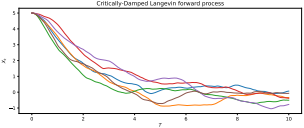
## Critically damped Langevin noising

$$\mathrm{d}\begin{pmatrix} \overrightarrow{X}_t \\ \overrightarrow{V}_t \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & \mathrm{I}_d \\ -\mathrm{I}_d & -2\mathrm{I}_d \end{pmatrix}}_{A} \begin{pmatrix} \overrightarrow{X}_t \\ \overrightarrow{V}_t \end{pmatrix} \mathrm{d}t + \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & 2\mathrm{I}_d \end{pmatrix}}_{\Sigma} \mathrm{d}B_t.$$

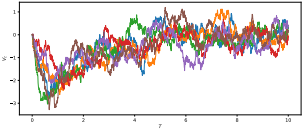
**Noising process comparison**




Standard OU noising




CLD position marginal $X_t$



CLD velocity marginal $V_t$

# What makes a forward SDE a generative model?

## 1. Interpolation

It transforms data into an easy-to-sample prior:

$$\pi_{\text{data}} \longrightarrow \pi_\infty .$$

## 2. Learnability

The reverse drift should be learnable from samples *i.e.*

$$\nabla \log p_t$$

should be learnable.

## 3. Sampling

The reverse dynamics should be numerically simulable.

## CLD forward process

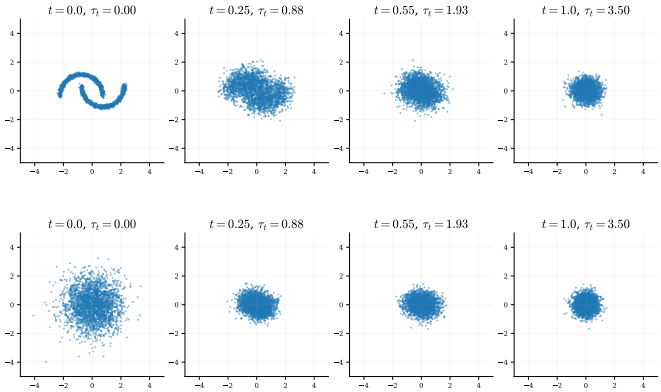
$$\mathrm{d}\overrightarrow{\mathbf{U}}_t = A\overrightarrow{\mathbf{U}}_t \mathrm{d}t + \Sigma \mathrm{d}B_t\,, \qquad \overrightarrow{\mathbf{U}}_0 \sim \pi_{\text{data}} \otimes \mathcal{N}(0, \mathrm{I}_d)\,.$$

$$\overrightarrow{\mathbf{U}}_t = \begin{pmatrix} \overrightarrow{X}_t \\ \overrightarrow{V}_t \end{pmatrix}, \quad A = \begin{pmatrix} 0 & \mathrm{I}_d \\ -\mathrm{I}_d & -2\mathrm{I}_d \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0 & 0 \\ 0 & 2\mathrm{I}_d \end{pmatrix}.$$

# 1. Interpolate between the data distribution and a prior

- The dynamics is linear on the extended space:

$$\overrightarrow{\mathbf{U}}_t = \mathrm{e}^{tA}\overrightarrow{\mathbf{U}}_0 + \int_0^t \mathrm{e}^{(t-s)A}\Sigma\,\mathrm{d}B_s, \qquad p_T \approx \pi_\infty = \mathcal{N}(0, \mathrm{I}_{2d}) \quad \text{for large } T.$$



# 2. The reverse process can be learned

## Time reversal on phase space

Let $p_t$ be the density of $\overrightarrow{\mathbf{U}}_t = (\overrightarrow{X}_t, \overrightarrow{V}_t) \in \mathbb{R}^{2d}$. Time reversal gives

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

and the reverse dynamics is

$$\mathrm{d}\overleftarrow{\mathbf{U}}_t = \left(-A\overleftarrow{\mathbf{U}}_t + \Sigma\Sigma^\top \nabla \log p_{T-t}(\overleftarrow{\mathbf{U}}_t)\right)\mathrm{d}t + \Sigma\,\mathrm{d}B_t.$$

## Learning the phase-space score

As in standard SGMs, train a score network on corrupted samples:

$$\mathcal{L}_{\mathrm{DSM}}(\theta) = \mathbb{E}\left[\left\|s_\theta(t, \overrightarrow{\mathbf{U}}_t) - \nabla \log p_t(\overrightarrow{\mathbf{U}}_t \mid \overrightarrow{\mathbf{U}}_0)\right\|^2\right].$$

Since $V_0 \sim \mathcal{N}(0, \mathrm{I}_d)$ is known, we may marginalize over $V_0$ and use the closed-form target

$$\mathcal{L}_{\mathrm{HSM}}(\theta) = \mathbb{E}\left[\left\|s_\theta(t, \overrightarrow{\mathbf{U}}_t) - \nabla \log p_t(\overrightarrow{\mathbf{U}}_t \mid \overrightarrow{X}_0)\right\|^2\right].$$

# 3. The reverse dynamics can be sampled

## Numerical simulation

- Euler–Maruyama baseline.
- Splitting / kinetic integrators.

## Takeaway

CLD keeps the SGM recipe, but changes the geometry of the reverse sampler. Empirically, this can improve generation quality.

Table 1: Unconditional CIFAR-10 generative performance.

| Class | Model | NLL↓ | FID↓ |
|---|---|---|---|
| **Score** | *CLD-SGM (Prob. Flow) (ours)* | ≤3.31 | 2.25 |
| | CLD-SGM (SDE) *(ours)* | - | 2.23 |
| **Score** | DDPM++, VPSDE (Prob. Flow) (Song et al., 2021c) | 3.13 | 3.08 |
| | DDPM++, VPSDE (SDE) (Song et al., 2021c) | - | 2.41 |
| | DDPM++, sub-VP (Prob. Flow) (Song et al., 2021c) | 2.99 | 2.92 |
| | DDPM++, sub-VP (SDE) (Song et al., 2021c) | - | 2.41 |
| | NCSN++, VESDE (SDE) (Song et al., 2021c) | - | 2.20 |
| | LSGM (Vahdat et al., 2021) | ≤3.43 | 2.10 |
| | LSGM-100M (Vahdat et al., 2021) | ≤2.96 | 4.60 |
| | DDPM (Ho et al., 2020) | ≤3.75 | 3.17 |
| | NCSN (Song & Ermon, 2019) | - | 25.3 |
| | Adversarial DSM (Jolicoeur-Martineau et al., 2021b) | - | 6.10 |
| | Likelihood SDE (Song et al., 2021b) | 2.84 | 2.87 |
| | DDIM (100 steps) (Song et al., 2021a) | - | 4.16 |
| | FastDDPM (100 steps) (Kong & Ping, 2021) | - | 2.86 |
| | Improved DDPM (Nichol & Dhariwal, 2021) | 3.37 | 2.90 |
| | VDM (Kingma et al., 2021) | ≤2.49 | 7.41 (4.00) |
| | UDM (Kim et al., 2021) | 3.04 | 2.33 |
| | D3PM (Austin et al., 2021) | ≤3.44 | 7.34 |
| | Gotta Go Fast (Jolicoeur-Martineau et al., 2021a) | - | 2.44 |
| | DDPM Distillation (Luhman & Luhman, 2021) | - | 9.36 |

Empirical comparison from (Dockhorn, 2022).

# Classical $\mathcal{W}_2$ contraction: reverse O.U. I

For the O.U. forward noising process, the reverse dynamics is

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

## Synchronous coupling

Fix $x, y \in \mathbb{R}^d$. Let $\overleftarrow{X}_t^x$ and $\overleftarrow{X}_t^y$ solve the same reverse SDE, driven by the same Brownian motion, with

$$\overleftarrow{X}_0^x = x, \qquad \overleftarrow{X}_0^y = y.$$

Set

$$Z_t := \overleftarrow{X}_t^x - \overleftarrow{X}_t^y.$$

The Brownian noises cancel, hence

$$\frac{\mathrm{d}}{\mathrm{d}t} Z_t = Z_t + 2\nabla \log p_{T-t}(\overleftarrow{X}_t^x) - \nabla \log p_{T-t}(\overleftarrow{X}_t^y).$$

Therefore

$$\frac{\mathrm{d}}{\mathrm{d}t}\|Z_t\|^2 = 2\|Z_t\|^2 + 4\left\langle Z_t, \nabla \log p_{T-t}(\overleftarrow{X}_t^x) - \nabla \log p_{T-t}(\overleftarrow{X}_t^y)\right\rangle.$$

# Classical $\mathcal{W}_2$ contraction: reverse O.U. II

## Uniform log-concavity assumption

Assume that $p_{T-t}$ is uniformly $\lambda$-log-concave:

$$\langle x-y, \nabla \log p_{T-t}(x) - \nabla \log p_{T-t}(y)\rangle \le -\lambda \|x-y\|^2.$$

Equivalently, if $p_{T-t}$ is siffuciently regular,

$$\nabla^2 \log p_{T-t} \preceq -\lambda \mathrm{I}_d.$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|Z_t\|^2 \le 2(1-2\lambda)\|Z_t\|^2.$$

By Grönwall,

$$\|Z_t\|^2 \le \exp\{-2(2\lambda-1)t\}\|Z_0\|^2.$$

## Consequence

If $\lambda > 1/2$, the reverse O.U. dynamics is contractive:

$$\mathcal{W}_2(\mu Q_t, \nu Q_t) \le \exp\{-(2\lambda-1)t\}\mathcal{W}_2(\mu,\nu),$$

where $Q_t$ is the reverse Markov kernel.

# Why the naive Euclidean contraction proof breaks for CLD

Consider two synchronously coupled reverse CLD processes and set $\mathbf{Z}_t = \overleftarrow{\mathbf{U}}_t^x - \overleftarrow{\mathbf{U}}_t^y$. For the reverse dynamics

$$\mathrm{d}\overleftarrow{\mathbf{U}}_t = \left(-A\overleftarrow{\mathbf{U}}_t + \Sigma\Sigma^\top \nabla \log p_{T-t}(\overleftarrow{\mathbf{U}}_t)\right)\mathrm{d}t + \Sigma\,\mathrm{d}B_t,$$

the Brownian noises cancel. By the mean-value theorem,

$$\nabla \log p_{T-t}(\overleftarrow{\mathbf{U}}_t^x) - \nabla \log p_{T-t}(\overleftarrow{\mathbf{U}}_t^y) = \bar{H}_t \mathbf{Z}_t,$$

with $\bar{H}_t$ an averaged Hessian of $\log p_{T-t}$. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{Z}_t\|^2 = -2\mathbf{Z}_t^\top A\mathbf{Z}_t + 2\mathbf{Z}_t^\top \Sigma\Sigma^\top \bar{H}_t \mathbf{Z}_t.$$

## Hypoelliptic obstruction

Even if $p_{T-t}$ is log-concave, so that $\bar{H}_t \preceq 0$, the term $\mathbf{Z}_t^\top \Sigma\Sigma^\top \bar{H}_t \mathbf{Z}_t$, is not necessarily nonpositive. Indeed,

$$\Sigma\Sigma^\top \bar{H}_t = \begin{pmatrix} 0 & 0 \\ 0 & 4\mathrm{I}_d \end{pmatrix} \begin{pmatrix} H_{xx} & H_{xv} \\ H_{vx} & H_{vv} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 4H_{vx} & 4H_{vv} \end{pmatrix}$$

only acts on the velocity directions.

⚠ Euclidean synchronous contraction is not automatic in hypoelliptic phase space.

# Solution: Long-term regularity of the renormalized score

**Idea.** Introduce a *renormalized* formulation of the backward process:

$$\mathrm{d}\overleftarrow{\mathbf{U}}_t = \tilde{A}\,\overleftarrow{\mathbf{U}}_t \mathrm{d}t + \Sigma\Sigma^\top \nabla \log \tilde{p}_{T-t}(\overleftarrow{\mathbf{U}}_t)\mathrm{d}t + \Sigma \mathrm{d}B_t, \qquad \tilde{p}_t := \frac{p_t}{p_\infty}.$$

**Key properties.**

1. $\tilde{A}$ is negative definite.
2. $\tilde{p}_t$ "quantifies" **deviation from equilibrium** $p_\infty$.
3. Its curvature $\nabla^2 \log \tilde{p}_t$ characterizes the **regularity of the score**, for all $t \in (0, T]$,

$$\left\|\nabla^2 \log \tilde{p}_t(\cdot)\right\| \;\leq\; C\left(1 + \tfrac{1}{\sqrt{t}}\right)\mathrm{e}^{-2at} \;=\; \tilde{L}_t.$$

4. Recover a bound of the type, as for general SGMs

$$\mathcal{W}_2(\pi_{\mathrm{data}}, \widehat{\pi}^{\theta}_{\infty,N}) \leq \mathrm{e}^{-T} c_1 + Mc_2 + \sqrt{h}c_3\,,$$

# Kinetic contraction mechanism

## Regularized score estimate

Under structural assumptions on the data distribution weaker than log-concavity, we proved

$$\|\nabla^2 \log \widetilde{p}_t\| \le \widetilde{L}_t, \qquad \widetilde{L}_t \lesssim \left(1 + \frac{1}{\sqrt{t}}\right) \mathrm{e}^{-2t}.$$

The singularity at $t = 0$ is integrable and the large-time behavior decays exponentially.

## Weighted contraction

There exist a positive definite matrix $\mathfrak{M}$ and $\eta > 0$ such that, for the synchronous difference,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \|\mathbf{Z}_t\|_{\mathfrak{M}}^2 &\le 2\mathbf{Z}_t^\top \mathfrak{M}\tilde{A}\mathbf{Z}_t + 2\mathbf{Z}_t^\top \mathfrak{M}\Sigma^2 \left(\nabla \log \tilde{p}_{T-t}\left(\overleftarrow{\mathbf{U}}_t^x\right) - \nabla \log \tilde{p}_{T-t}\left(\overleftarrow{\mathbf{U}}_t^y\right)\right) \\ &\le 2(-\eta + \sigma^2 \tilde{L}_t) \|\mathbf{Z}_t\|_{\mathfrak{M}}^2 \ .\end{aligned}$$

Hence, by Grönwall, there exists $C \ge 0$ such that

$$\begin{aligned}\|\mathbf{Z}_t\|_{\mathfrak{M}}^2 &\le \mathrm{e}^{-2\eta t + \sigma^2 \int_0^t \tilde{L}_s \mathrm{d}s} \|\mathbf{Z}_0\|_{\mathfrak{M}}^2 \\ &\le C\mathrm{e}^{-2\eta t} \|\mathbf{Z}_0\|_{\mathfrak{M}}^2 \ ,\end{aligned}$$

# Selected references


- 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*.
- Song and Ermon (2019). *Generative modeling by estimating gradients of the data distribution*.
- Song et al. (2021). *Score-based generative modeling through stochastic differential equations*.
- Dockhorn et al. (2022). *Score-based generative modeling with critically-damped Langevin diffusion*.
- Strasman, Surendran, Boyer, Le Corff, Lemaire and Ocello (2025). *Wasserstein stability of kinetic score-based generative models*.

# Solution 2: restore ellipticity

## Idea

Inject a small amount of noise on *all* coordinates:

$$\Sigma_\varepsilon = \begin{pmatrix} \varepsilon & 0 \\ 0 & \sigma \end{pmatrix} \otimes \mathrm{I}_d, \qquad \varepsilon > 0.$$

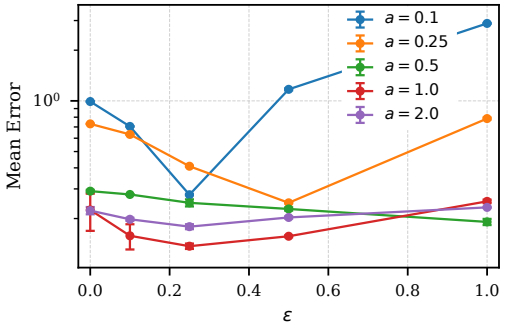
## Consequences

- **Uniform ellipticity**: full phase-space noise.
- **More quantitative bounds**: standard log-concave tools apply.
- **Practice**: $\varepsilon$ is an extra parameter controlling path regularity.

# Numerical aspects

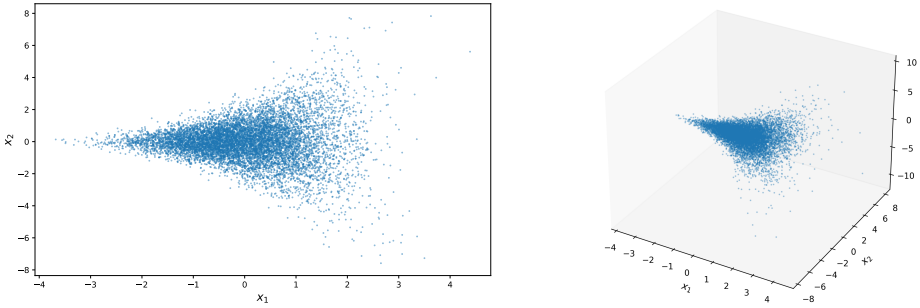
## Empirics: Funnel dataset, $d = 100$

- Small $\varepsilon > 0$ can improve sliced-$\mathcal{W}_2$ compared with the CLD baseline ($\varepsilon = 0$).
- There is a trade-off with the other model hyperparameters.




Mean sliced-$\mathcal{W}_2$ over 5 runs; error bars are $\pm 1$ standard deviation.

# Funnel distribution




Samples from a funnel distribution. Left: first two coordinates. Right: first three coordinates.