SCIENCES SORBONNE UNIVERSITÉ

Critically-Damped Langevin Score-based Generative Models:

introduction, motivation and convergence.

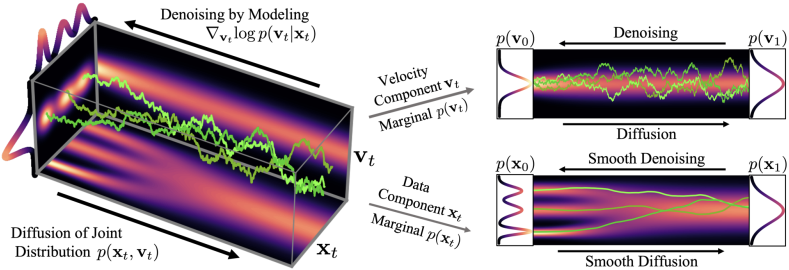
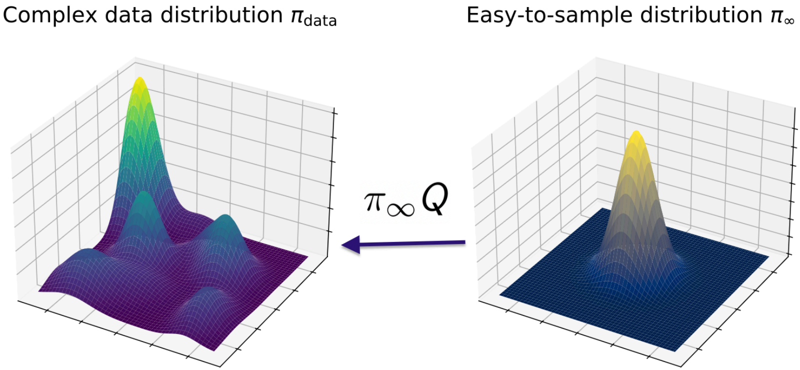


Image taken from Dockhorn et al. (2022).

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 π_{data}
- Goal: **generate new samples from** π_{data} (i.e. find a proba π_∞ and a simulable kernel Q such that $\pi_{\mathrm{data}} \simeq \pi_\infty Q$).



SGMs Philosophy.

- "Creating noise from data is easy; creating data from noise is generative modeling." (Song et al., 2021)

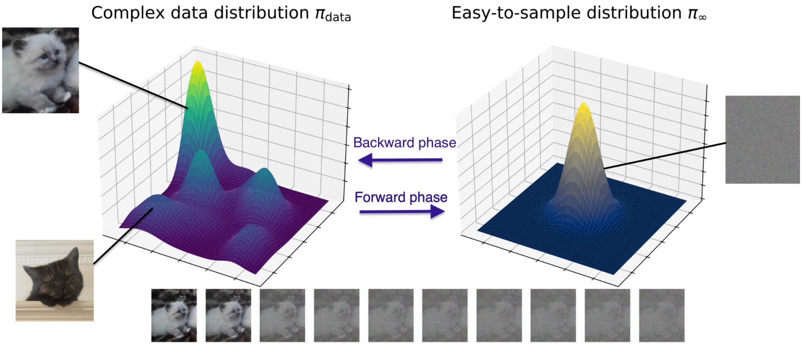


Table of Contents

Table of Contents

From data to noise: the forward process.

- $\left(\overrightarrow{X}_t\right)_{t\in[0,T]}$ is solution to an **Ornstein–Uhlenbeck** process:

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

- If unfamiliar with SDEs: limit of a discrete-time process given by

$$X_{k+h} = \sqrt{1-2h}\,X_k + \sqrt{2h}\,Z_k, \quad Z_k \sim \mathcal{N}(0, I_d), \quad h \to 0\,.$$

- **Intuition:** destroys signal via Gaussian noise and rescaling.

SGMs through SDE: more on the forward process.

- The noising procedure implies a scaling down of the data points $\mathrm{d}\overrightarrow{X}_t = -\overrightarrow{X}_t \mathrm{d}t$,

SGMs through SDE: more on the forward process.

- ... and a Gaussian noising process $\mathrm{d}\overrightarrow{X}_t = \sqrt{2}\mathrm{d}B_t$,

SGMs through SDE: more on the forward process.

From noise to data: the backward process.

- This forward process admits a **time-reversed process** (Anderson, 1982; Cattiaux et al., 2021), i.e.

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

with,

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

with p_t the p.d.f. of $\overrightarrow{X}_t$.

- The score term drives the backward process towards **regions of high probability**.
- This is (**almost**) a **generative model**: $\overleftarrow{X}_T \sim \pi_{\mathrm{data}}$.

Table of Contents

SGMs in Practice I: mixing time.

- Let Q_t be the semigroup of $\overleftarrow{X}_t$:

$$Q_t(x, \mathrm{d}y) = \mathbb{P}\left(\overleftarrow{X}_t \in \mathrm{d}y | \overleftarrow{X}_0 = x\right) .$$

- **Time-reversal holds when $\overleftarrow{X}_\mathbf{0} \sim \mathbf{p_T}$**, i.e.

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

- But p_t depends on π_{data}:

$$p_t(x_t) = \int_{\mathbb{R}^d} \underbrace{p_t(x_t|x_0)}_{\text{p.d.f. of } \overrightarrow{X}_t | \overrightarrow{X}_0} \pi_{\mathsf{data}}(\mathrm{d}x_0) .$$

- I practice, one wants an **independent** and **easy-to-sample** probability π_∞ to initialize the generative model.

SGMs in Practice I: mixing time.

- 💡 leverage the ergodicity of the O–U kernel.
- **Forward process** admits time marginal with $Z \sim \mathcal{N}(0, I_d)$ and $Z \perp X_0$:

$$\overrightarrow{X}_t = \mathrm{e}^{-t}\overrightarrow{X}_0 + \sqrt{1 - \mathrm{e}^{-2t}}Z$$

- For T large, the initial conditions are forgotten:

$$p_T \approx \pi_\infty \sim \mathcal{N}\left(0, I_d\right) .$$

⚠ **Mixing Time Error:** $\pi_{\text{data}} \simeq \pi_\infty Q_T$

SGMs in practice II: learn the score function.

- The backward process depends on the score function $\nabla \log p_t(x)$.
- The forward process marginals can be sampled exactly.
- Train a **deep neural network** $s_\theta : [0, T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ to minimize:

$$\mathcal{L}_{\mathrm{naive}}(\theta) = \mathbb{E}\left[\left\| s_\theta\left(\tau, \overrightarrow{X}_\tau\right) - \nabla \log p_\tau\left(\overrightarrow{X}_\tau\right)\right\|^2\right] ,$$

  with $\tau \sim \mathcal{U}(0, T)$ independent of the forward process $(\overrightarrow{X}_t)_{t \geq 0}$.
- But $p_\tau(x)$ is **unknown !**

SGMs in practice II: learn the score function.

- 💡 its **conditional version** shares the same optimum (Hyvärinen and Dayan, 2005; Vincent, 2011):

$$\mathcal{L}_{\mathrm{score}}(\theta) = \mathbb{E}\left[\| s_\theta \left(\tau, \overrightarrow{X}_\tau\right) - \nabla \log p_\tau \left(\overrightarrow{X}_\tau | \overrightarrow{X}_0\right) \|^2 \right] .$$

- The conditional score is explicit:

$$\nabla \log p_\tau(\overrightarrow{X}_\tau | \overrightarrow{X}_0) = \frac{m_\tau \overrightarrow{X}_0 - \overrightarrow{X}_\tau}{\sigma_{\tau^2}} = -\frac{Z}{\sigma_\tau}$$

  with $m_\tau = \mathrm{e}^{-\tau}$ and $\sigma_\tau = \sqrt{1 - m_\tau^2}$.
- Score matching Neural Networks writes as,

$$\mathcal{L}_{\mathrm{score}}(\theta) = \mathbb{E}\left[\left\| s_\theta \left(\tau, \overrightarrow{X}_\tau\right) + \frac{Z}{\sigma_\tau} \right\|^2 \right] .$$

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

SGMs in practice III: simulate from the backward kernel.

- The backward drift is **non-linear**: non-Gaussian.
- 💡 discretize $[0, T]$ in N steps with $t_k = kh$, $h = T/N$.
- Euler–Maruyama discretization:

$$\bar{X}_{t_{k+1}} = \bar{X}_{t_k} + h\Big(\bar{X}_{t_k} + 2\, s_\theta(T - t_k, \bar{X}_{t_k})\Big) + \sqrt{2h}\, Z_k$$

- Other approaches preserving the time marginals exist (*e.g.* ODE sampling).

⚠ **Discretization error**: $\pi_{\mathrm{data}} \approx \pi_\infty Q^\theta_{T,N} := \widehat{\pi}^\theta_{\infty,N}$

Table of Contents

Draw inspiration from MCMC.

- In **sampling,** one wants to sample from $\pi \propto \mathrm{e}^{-U}$.
  - When $U : \mathbb{R}^d \to \mathbb{R}$ is smooth and typically strongly convex,

$$\mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

    admits π as invariant measure.
  - Sampling can be done by discretization (ULA) or accept–reject corrections (MALA).
- This can be extended to a **kinetic setting**:

$$\mathrm{d}\begin{pmatrix} X_t \\ V_t \end{pmatrix} = \begin{pmatrix} V_t \\ -(V_t + \nabla U(X_t)) \end{pmatrix} \mathrm{d}t + \sqrt{2}\begin{pmatrix} 0 \\ 1 \end{pmatrix} \mathrm{d}B_t$$

- Stationary distribution $\pi(\mathrm{d}x, \mathrm{d}v) \propto \mathrm{e}^{-U(x) - \frac{\|v\|^2}{2}} \mathrm{d}x\mathrm{d}v$.

Extending the Phase Space of SGMs.

- Augment data space with a **velocity component** $(V_t)_{t\in[0,T]}$.
- X_t and V_t are coupled through *Hamiltonian-like* interactions.
- **Noise injection** only on the **velocity** component.

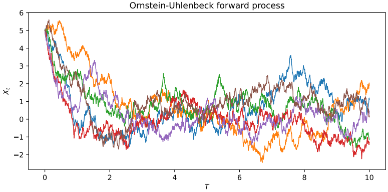
Forward process: for $\overrightarrow{\mathbf{U}}_t = (\overrightarrow{X}_t, \overrightarrow{V}_t)^\top \in \mathbb{R}^2$ and $B_t \in \mathbb{R}^2$,

$$\mathrm{d}\begin{pmatrix} X_t \\ V_t \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & -2 \end{pmatrix} \begin{pmatrix} X_t \\ V_t \end{pmatrix} \mathrm{d}t + \begin{pmatrix} 0 & 0 \\ 0 & \sigma \end{pmatrix} \mathrm{d}B_t \, , \; (X_0, V_0) \sim \pi_{\mathrm{data}} \otimes \pi_v \, ,$$

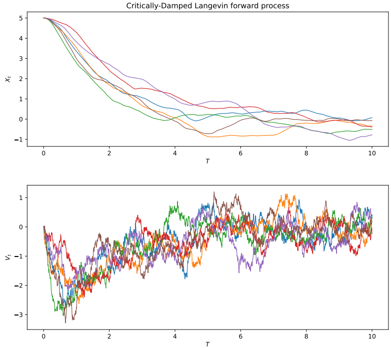
where $\pi_v \sim \mathcal{N}(0, v^2)$. We use compact matrix notation,

$$\mathrm{d}\overrightarrow{\mathbf{U}}_t = A\overrightarrow{\mathbf{U}}_t \mathrm{d}t + \Sigma \mathrm{d}B_t \, , \quad \overrightarrow{\mathbf{U}}_0 \sim \pi_{\mathrm{data}} \otimes \pi_v \, .$$

Noising process comparison.



OU process



CLD process (position & velocity)

Table of Contents

What makes a forward SDE a generative model?

- A well-chosen **noising process** must satisfy three properties:
  1. **Interpolation:** transforms the data distribution π_{data} into an easy-to-sample prior π_{∞}.
  2. **Learnability:** time-dependent score functions $\nabla \log p_t(\cdot)$ can be learned.
  3. **Efficient sampling:** marginals can be efficiently simulated.
- Examples: Variance-Preserving (VP), Variance-Exploding (VE), flow matching and **Critically-Damped Langevin (CLD) diffusions**.

1. Interpolate between the data distribution and a prior.

- The forward process evolves in the **extended phase space** $\overrightarrow{\mathbf{U}}_t = (\overrightarrow{X}_t, \overrightarrow{V}_t)^\top \in \mathbb{R}^2$ as

$$\overrightarrow{\mathbf{U}}_t = \mathrm{e}^{tA}\overrightarrow{\mathbf{U}}_0 + \int_0^t \mathrm{e}^{(t-s)A}\Sigma \mathrm{d}B_s \,. \qquad (1)$$

  and converges to $\pi_\infty \sim \mathcal{N}(0_2, \Sigma_\infty)$.
- Time reversal property applies on the extended space, *i.e.*

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

- leading to the backward SDE:

$$\mathrm{d}\overleftarrow{\mathbf{U}}_t = -A\overleftarrow{\mathbf{U}}_t \mathrm{d}t + \Sigma^2 \nabla \log p_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)\mathrm{d}t + \Sigma \mathrm{d}B_t \,,$$

  with $p_t(x, v)$ de p.d.f of (1).

2. Score function can be learned

- As before s_θ trained to learn conditional score but on the whole phase-space state $\overrightarrow{\mathbf{U}}_t = (\overrightarrow{X}_t, \overrightarrow{V}_t)$:

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

- However, we know that $\overrightarrow{V}_0 \sim \mathcal{N}(0, v^2)$, so we can marginalize $\overrightarrow{\mathbf{U}}_0 = (\overrightarrow{X}_0, \overrightarrow{V}_0)^\top$ over V_0, leading to a **closed-form expression** of $\nabla \log p_t(\overrightarrow{\mathbf{U}}_t \mid \overrightarrow{X}_0)$:

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

yielding **more stable training** objective.

3. Marginals can be sampled.

Different numerical schemes:

- Euler–Maruyama (standard baseline);
- Symplectic integrators design for position-velocity state-spaces.

Combining all this leads to **better numerical performance** (Dockhorn et al., 2022):

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

Table of Contents

A variety of convergence results

- Assume score function is appropriately learned *e.g.*

$$\|s_\theta(t, \mathbf{U}_t) - \nabla \log p_t(\mathbf{U}_t)\|_{L_2} \leq M$$

  where the expectation is taken under some appropriately chosen stochastic process.
- Using this framework a **variety of upper bounds** to the **distance between the data distribution and the generated distribution** $d\left(\pi_{\mathrm{data}}, \hat{\pi}\right)$ have been established for various metrics:
  - For the total variation distance and Kullback-Leibler divergence: De Bortoli et al. (2021); Conforti et al. (2023); Bortoli et al. (2023); Chen et al. (2023); Chen (2023).
  - For the Wasserstein distance: Lee et al. (2022, 2023); Bruno et al. (2023); Gao et al. (2023).

Wasserstein–2: upper bounds.

- The $\mathcal{W}_2$ distance is defined as

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

- Control the errors already presented:

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

$$+ \underbrace{\mathcal{W}_2(\mathcal{L}(\bar{X}_{\infty,N}), \mathcal{L}(\bar{X}^{\theta}_{\infty,N})))}_{\textit{Score approx.}}$$

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

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

Backward contraction for O.U forward I.

- Proof relies **mostly** on **contraction for Euclidean norm**:

$$\mathcal{W}_2^2(\pi_{\mathrm{data}}, \widehat{\pi}^{\theta}_{\infty,N}) \leq \left\| \overrightarrow{X}_0 - \bar{X}^{\theta}_{\infty,N} \right\|^2_{L_2}$$

- Fix $x, y \in (\mathbb{R})^2$:

$$\mathrm{d}\overleftarrow{X}^x_t = \left(\overleftarrow{X}^x_t + 2\nabla \log p_{T-t}\left(\overleftarrow{X}^x_t\right)\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \quad \overleftarrow{X}_0 = x \; p.s.$$

$$\mathrm{d}\overleftarrow{X}^y_t = \left(\overleftarrow{X}^y_t + 2\nabla \log p_{T-t}\left(\overleftarrow{X}^y_t\right)\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \quad \overleftarrow{X}_0 = y \; p.s.$$

- Consider a synchronous coupling and introduce the difference ODE $Z_t = \overleftarrow{X}^x_t - \overleftarrow{X}^y_t$, which satisfies

$$\mathrm{d}Z_t = Z_t + 2\underbrace{\left(\nabla \log p_{T-t}(\overleftarrow{X}^x_t) - \nabla \log p_{T-t}(\overleftarrow{X}^y_t)\right)}_{:=\Delta_t}\mathrm{d}t$$

- and study

$$\frac{\mathrm{d}}{\mathrm{d}t}\|Z_t\|^2 = 2Z_t^\top(\mathrm{d}Z_t) = 2\left(\|Z_t\|^2 + 2\langle Z_t, \Delta_t\rangle\right).$$

Backward contraction for O.U forward II.

- If p_{T-t} is λ–**log-concave**, then, there exists $\lambda > 0$, such that
  - $\langle Z_t, \nabla \log p_{T-t}(\overleftarrow{X}_t^x) - \nabla \log p_{T-t}(\overleftarrow{X}_t^y) \rangle \leq -\lambda \|Z_t\|^2$;
  - $\nabla^2 \log p_{T-t} \preccurlyeq -\lambda I_d$.
- Therefore, using Grönwall inequality,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\|Z_t\|^2 &\leq 2(1-2\lambda)\|Z_t\|^2 \\ &\leq \mathrm{e}^{2(1-2\lambda)t}\|Z_0\|^2. \end{aligned}$$

- **Takeaway:** strong log-concavity ($\lambda > 1/2$) **gives contraction** for $\| \cdot \|$ which implies $\mathcal{W}_2$ contraction.

Table of Contents

Log-concavity is not enough for CLD

$$\mathrm{d}\overleftarrow{\mathbf{U}}_t = -A\overleftarrow{\mathbf{U}}_t\mathrm{d}t + \Sigma^2\nabla \log p_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)\mathrm{d}t + \Sigma\mathrm{d}B_t\,,$$

- Consider a synchronous coupling and study the stability of the difference ODE $\mathbf{Z}_t = \overleftarrow{\mathbf{U}}^x_t - \overleftarrow{\mathbf{U}}^y_t$, which satisfies
$$\frac{\mathrm{d}}{\mathrm{d}t}(\|\mathbf{Z}_t\|^2) = -2\mathbf{Z}_t^\top A\mathbf{Z}_t + 2\,\mathbf{Z}_t^\top\Sigma^2 H_t\mathbf{Z}_t.$$

with

$$\Sigma = \begin{pmatrix} 0 & 0 \\ 0 & \sigma \end{pmatrix}$$

where we used the mean value theorem with

$$H_t = \nabla^2 logp_{T-t} = \begin{pmatrix} H_t^{xx} & H_t^{xv} \\ H_{t^x v} & H_t^{vv} \end{pmatrix}.$$

Log-concavity is not enough for CLD (counterexample).

Take $p_t(x, v)$ that is $(1 - c)$-**log-concave** with,

$$H_t = \begin{pmatrix} -1 & c \\ c & -1 \end{pmatrix}, \quad 0 < c < 1; \quad \mathrm{Spec}(H_t) = -1 \pm c\,,$$

Then with $\Sigma = \mathrm{diag}(0, \sigma)$ and $\mathbf{Z} = (Z_x, Z_v)$,

$$2\,\mathbf{Z}^\top \Sigma^2 H_t \mathbf{Z} = 2\sigma^2\big(-c\,Z_x Z_v - Z_v^2\big).$$

Choose $Z_x > 0$, $Z_v < 0$ with ratio $|Z_x|/|Z_v| > 1/c$. Then the RHS becomes **positive**.

Takeaway: even though p_t is log-concave, the projected curvature $\Sigma^2 H_t$ is *not* negative semidefinite. Uniform contraction is hopeless.

Solution 1: Long-term regularity of the renormalized score

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

$$\mathrm{d}\overleftarrow{\mathbf{U}}_t = \tilde{A}\,\overleftarrow{\mathbf{U}}_t\mathrm{d}t + \Sigma^2\nabla\log\tilde{p}_{T-t}(\overleftarrow{\mathbf{U}}_t)\mathrm{d}t + \Sigma\mathrm{d}B_t, \qquad \tilde{p}_t := \frac{p_t}{p_\infty}.$$

Key properties.

1. $\tilde{A}$ is **negative definite**.
2. $\tilde{p}_t$ "quantifies" **deviation from equilibrium** p_∞.
3. Its curvature $\nabla^2\log\tilde{p}_t$ characterizes the **regularity of the score**.

Structure & regularity assumptions on p_{data}

Finite relative Fisher information

$$\mathcal{I}(p_{\mathrm{data}} \mid p_\infty) = \int_{\mathbb{R}^d} \left\| \nabla \log \frac{p_{\mathrm{data}}}{p_\infty}(x) \right\|^2 p_{\mathrm{data}}(x) \mathrm{d}x < \infty \,.$$

Log-Lipschitz perturbation of a strongly log-concave base

$$p_{\mathrm{data}}(x) \propto \exp\left(-[V(x) + H(x)] \right),$$

with for all $x, y \in (\mathbb{R}^d)^2$:

- $\exists \alpha > 0$ such that $\alpha I_d \preceq \nabla^2 V(x)$;
- $|H(x) - H(y)| \leq L\|x - y\|$.

One–sided Lipschitz score

$$-\left(\nabla \log p_{\mathrm{data}}(x) - \nabla \log p_{\mathrm{data}}(y)\right)^\top (x - y) \;\leq\; L_0\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

Regularity of the renormalized score

- Under the previous hypotheses, there exists a constant $C > 0$ such that, for all $t \in (0, T]$,

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

- **Interpretation.**
  - **Short times ($t \to 0$):** the singularity in $1/\sqrt{t}$ remains **integrable**.
  - **Long times ($t \to \infty$):** exponential decay.

- **Takeaway:** The renormalized score function **regularizes over time**.

Solution 1: Long-term regularity of the renormalized score

- Informally, there exists $\mathfrak{M}$ PSD matrix and $\eta > 0$ such that

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

Using Grönwall's lemma, there exists $C > 0$, such that,

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

Contraction!

Final $\mathcal{W}_2$ upper bound

- From contraction in the extended phase space, the three sources of errors (mixing, approximation, and discretization) can be controlled jointly:

$$\mathcal{W}_2\Big(\pi_{\mathrm{data}} \otimes \pi_v,\, \mathcal{L}\big(\bar{\mathbf{U}}_T^\theta\big)\Big) \le c_1\, \mathrm{e}^{-c_2 T}\, \mathcal{W}_2(\pi_{\mathrm{data}} \otimes \pi_v,\, \pi_\infty) + c_1 \sigma^2 M + c_1 \sqrt{h}\,.$$

- Projecting onto the position component X ($P_X(x, v) = x$) preserves the $\mathcal{W}_2$ distance, since P_X is 1–Lipschitz:

$$\mathcal{W}_2\Big(\pi_{\mathrm{data}},\, \mathcal{L}\big(\bar{X}_T^\theta\big)\Big) \le \mathcal{W}_2\Big(\pi_{\mathrm{data}} \otimes \pi_v,\, \mathcal{L}\big(\bar{\mathbf{U}}_T^\theta\big)\Big)\,.$$

Solution 2: restore ellipticity

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

$$\Sigma = \begin{pmatrix} \varepsilon & 0 \\ 0 & \sigma \end{pmatrix}, \qquad \varepsilon > 0.$$

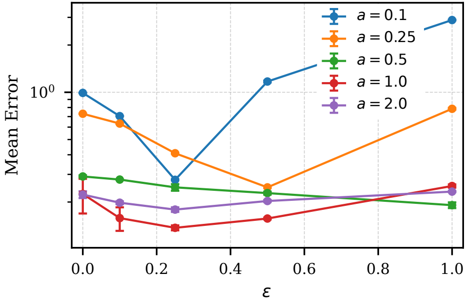
Consequences.

- **Uniform ellipticity**: (multi-dimensional O.U. structure).
- **More quantitative bounds :** standard log-concave tools apply.
- **Practice:** ε provides a new parameters to control the regularity of the sample paths.

Solution 2: Numerical aspects

Empirics (Funnel dataset, $d = 100$).

- Small ε often **improves** sliced-$\mathcal{W}_2$ vs. $\varepsilon = 0$ (CLD baseline).
- **Trade-off:** slight sensitivity to other hyperparameters.



Mean $\mathcal{W}_2$ over 5 runs; error bars represent $\pm$ one standard deviation.

Funnel distribution scatter plot

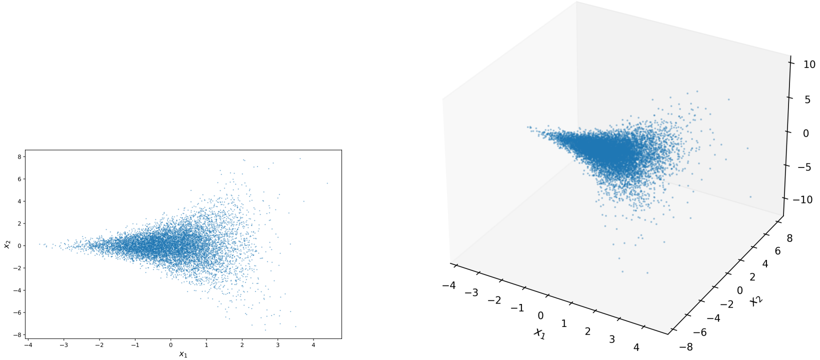
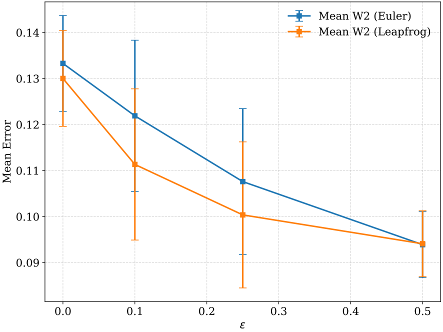


Figure: 10 000 samples from a funnel distribution in dimension 50. Plot of the 1st and 2nd dimension (left) and plot of the 1st, 2nd and 3rd dimension (right).

Solution 2: Numerical aspects

- Even with a small $\varepsilon > 0$, **structure-preserving integrators** can further improve performance.
- **But:** higher computational cost — the network must learn full gradients $\nabla \log p_t(x, v)$ instead of velocity-only terms $\nabla_v \log p_t(v)$, **doubling the effective dimension**.



ooo

B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

V. D. Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling, 2023.

S. Bruno, Y. Zhang, D.-Y. Lim, Ö. D. Akyildiz, and S. Sabanis. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *arXiv preprint arXiv:2311.13584*, 2023.

P. Cattiaux, G. Conforti, I. Gentil, and C. Léonard. Time reversal of diffusion processes under a finite entropy condition. *arXiv preprint arXiv:2104.07708*, 2021.

S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions, 2023.

T. Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.

G. Conforti, A. Durmus, and M. G. Silveri. Score diffusion models without early stopping: finite fisher information is all you need, 2023.

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

T. Dockhorn, A. Vahdat, and K. Kreis. Score-based generative modeling with critically-damped langevin diffusion. In *International Conference on Learning Representations (ICLR)*, 2022.

X. Gao, H. M. Nguyen, and L. Zhu. Wasserstein convergence guarantees for a general class of score-based generative models, 2023.

A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*, 2021.

P. Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.