Wasserstein Convergence of Critically Damped Langevin Diffusions

Stanislas Strasman[1], Sobihan Surendran[1,2], Claire Boyer[3], Sylvain Le Corff[1], Vincent Lemaire[1], Antonio Ocello[4]

[1]Sorbonne Université and Université Paris Cité, CNRS, LPSM, F-75005 Paris, France [2]LOPF, Califrais' Machine Learning Lab, Paris, France

[3]LMO, Université Paris-Saclay, UMR CNRS 8628, Institut Universitaire de France, Orsay, France

[4]CREST, Groupe ENSAE-ENSAI, ENSAE Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France

SCAN ME

Introduction

Consider $\mathcal{D} = \{x_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$ i.i.d. from an unknown distribution π_{data}.

Goal. Learn a generative mechanism whose output distribution $\widehat{\pi}$ is close to π_{data} (e.g., in Wasserstein distance $\mathcal{W}_2$).

Score-based Generative Models (SGMs). Construct

- a *forward noising process* transporting π_{data} to a simple prior π_∞,
- a *backward denoising process* mapping noise samples back to data.

Classical SGMs. Starting from $\overrightarrow{X}_0 \sim \pi_{\text{data}}$, common forward processes include

VP SDE: $\mathrm{d}\overrightarrow{X}_t = -\overrightarrow{X}_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t,$

VE SDE: $\mathrm{d}\overrightarrow{X}_t = \sqrt{2}\,\mathrm{d}B_t,$

Flow matching: $\overrightarrow{X}_t = (1-t)\overrightarrow{X}_0 + tZ, \quad Z \sim \mathcal{N}(0, I_d) \perp \overrightarrow{X}_0, \; t \in [0,1].$

Kinetic SGMs (this work) evolve in an extended position/velocity phase space.

Framework

Forward process. Kinetic phase-space dynamics for $\overrightarrow{\mathbf{U}}_t = (\overrightarrow{X}_t, \overrightarrow{V}_t) \in \mathbb{R}^{2d}$:

$$\mathrm{d}\overrightarrow{\mathbf{U}}_t = A\overrightarrow{\mathbf{U}}_t\mathrm{d}t + \Sigma\mathrm{d}B_t, \qquad \overrightarrow{\mathbf{U}}_0 \sim \pi_{\text{data}} \otimes \pi_v, \tag{1}$$

with

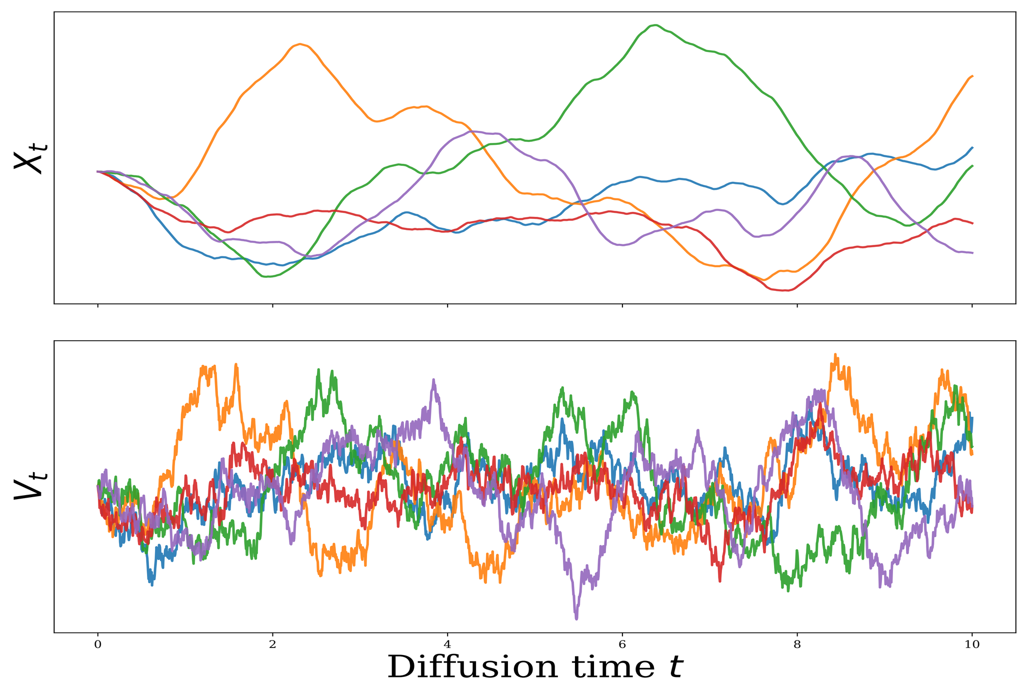
$$A = \begin{pmatrix} 0 & a^2 \\ -\mathbf{I}_d & -2a\mathbf{I}_d \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 0 & 0 \\ 0 & \sigma\mathbf{I}_d \end{pmatrix}, \qquad \pi_v = \mathcal{N}(0, v^2\mathbf{I}_d).$$

☞ *Hamiltonian-like coupling of (X, V), with noise injected only in the velocity.*

Backward process. [1] $(\overleftarrow{\mathbf{U}}_t)_{t\in[0,T]} \overset{\mathcal{L}}{=} (\overrightarrow{\mathbf{U}}_{T-t})_{t\in[0,T]}$ follows

$$\mathrm{d}\overleftarrow{\mathbf{U}}_t = -A\overleftarrow{\mathbf{U}}_t\,\mathrm{d}t + \Sigma^2\nabla\log p_{T-t}(\overleftarrow{\mathbf{U}}_t)\,\mathrm{d}t + \Sigma\,\mathrm{d}B_t, \qquad \overleftarrow{\mathbf{U}}_0 \sim p_T, \tag{2}$$

with p_t the p.d.f. of (1). Let Q_t be the semigroup of (2), so that $\pi_{\text{data}} = p_T Q_T$.



Prior work

- Empirical evidence that CLDs outperform standard SGMs in practice [2].
- KL convergence analyses of kinetic Langevin dynamics for the special case $a = 1$, $\sigma = 2$ [3, 4].
- **Gap:** no *Wasserstein* convergence guarantees are known for CLDs.

SGM-CLDs in Practice

☞ p_T is not accessible, but for large T, the process forgets its initialization.

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

■ **Mixing-time error:** $\pi_{\text{data}} \approx \pi_\infty Q_T$

☞ Score function $\nabla \log p_t$ is intractable but can be approximated by a deep neural network s_θ via score matching.

■ **Approximation error:** $\pi_{\text{data}} \approx \pi_\infty Q_T^\theta$

☞ Backward drift is **non-linear** and should be discretized into N finite steps.

■ **Discretization error:** $\pi_{\text{data}} \approx \pi_\infty Q_{T,N}^\theta := \widehat{\pi}_{\infty,N}^\theta$

Convergence results for SGMs rely on controlling each of the sources of error:

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

⚠ **Overcoming Hypoellipticity.** For non-kinetic SGMs, these terms are controlled by establishing a **contraction property in the Euclidean norm** via coupling arguments under **strong log-concavity** of p_t [5]. This condition no longer suffices for CLD: the dynamics are **hypoelliptic** (noise only in velocity).

Solution 1: long-term Lipschitz regularity of the renormalized score

Let p_∞ be the invariant density of the forward CLD and $\tilde{p}_t := p_t/p_\infty$. The **renormalized backward dynamics** writes as

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

where $\tilde{A}$ is a negative definite matrix.

Lipschitz continuity of renormalized score (exponential decay). Under regularity assumptions on π_{data}, there exists $C > 0$ such that, for all $t \in (0, T]$:

$$\|\nabla^2 \log \tilde{p}_t\| \;\le\; C\Big(1 + \tfrac{1}{\sqrt{t}}\Big)e^{-2at} \;=:\; \tilde{L}_t\,.$$

Norm contraction. $\tilde{L}_t$ is integrable and the backward flow contracts in a weighted norm, there exists $\eta > 0$:

$$\|\overleftarrow{\mathbf{U}}_t^x - \overleftarrow{\mathbf{U}}_t^y\|_{\mathfrak{M}} \;\le\; Ce^{-\eta t}\|\overleftarrow{\mathbf{U}}_0^x - \overleftarrow{\mathbf{U}}_0^y\|_{\mathfrak{M}}.$$

Wasserstein Convergence Analysis of CLD

Theorem: Under mild regularity assumptions on π_{data}, there exist constants $c_1, c_2, c_3 > 0$ such that

$$\mathcal{W}_2(\pi_{\text{data}}, \widehat{\pi}_{\infty,N}^\theta) \;\le\; c_1 e^{-c_2 T} \;+\; c_2 M \;+\; c_3\sqrt{T/N},$$

with

$$\sup_{k\in\{0,\dots,N-1\}} \left\|\nabla \log \tilde{p}_{T-t_k}(\bar{\mathbf{U}}_{t_k}^\theta) - s_\theta\left(T - t_k, \bar{\mathbf{U}}_{t_k}^\theta\right)\right\|_{L_2} \le M\,.$$

Solution 2: restore ellipticity

Add a small amount of noise in the position coordinates:

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

Consequences.

- ▶ **Uniform ellipticity**: matrix Ornstein-Uhlenbeck process.
- ▶ **Standard analysis restored**: log-concave contraction in the Euclidean metric, enabling sharper quantitative bounds.
- ▶ **Practical effect**: ε controls sample-path smoothness.

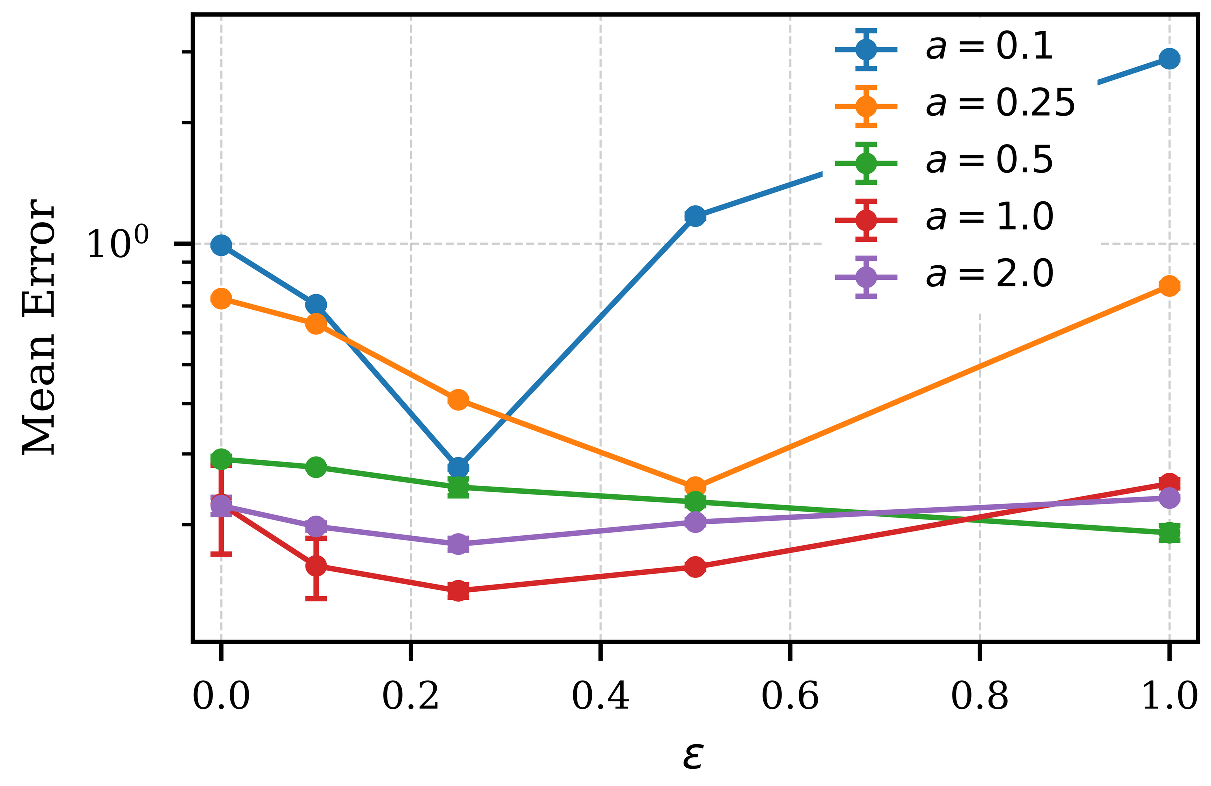


Figure: Mean sliced-$\mathcal{W}_2$ on the 100-dimensional **Funnel** distribution.

⇒ *Introducing a small regularization parameter ε improves generation quality.*

Additional remarks.

- Even with a small $\varepsilon > 0$, **structure-preserving integrators** can further improve performance.
- **Trade-off:** training becomes more expensive, since the network must learn full gradients $\nabla \log p_t(x, v)$ instead of velocity-only terms $\nabla_v \log p_t(v)$, effectively **doubling the dimension**.

References

[1] Patrick Cattiaux et al. "Time reversal of diffusion processes under a finite entropy condition". In: *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques* 59.4 (2023), pp. 1844–1881.

[2] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. "Score-Based Generative Modeling with Critically-Damped Langevin Diffusion". In: *International Conference on Learning Representations*. 2022.

[3] Sitan Chen et al. "Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions". In: *International Conference on Learning Representations*. 2023.

[4] Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. "KL convergence guarantees for score diffusion models under minimal data assumptions". In: *SIAM Journal on Mathematics of Data Science* 7.1 (2025), pp. 86–109.

[5] Xuefeng Gao, Hoang M Nguyen, and Lingjiong Zhu. "Wasserstein convergence guarantees for a general class of score-based generative models". In: *Journal of Machine Learning Research* 26.43 (2025), pp. 1–54.