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, ENSAE, Institut Polytechnique de Paris, Palaiseau, France

Abstract

Score-based Generative Models (SGMs) have achieved impressive performance in data generation across a wide range of applications and benefit from strong theoretical guarantees. Recently, methods inspired by statistical mechanics, in particular, Hamiltonian dynamics, have introduced Critically-damped Langevin Diffusions (CLDs), which define diffusion processes on extended spaces by coupling the data with auxiliary variables. These approaches, along with their associated score-matching and sampling procedures, have been shown to outperform standard diffusion-based samplers numerically. In this paper, we analyze a generalized dynamic that extends classical CLDs by introducing an additional hyperparameter controlling the noise applied to the data coordinate, thereby better exploiting the extended space. We further derive a novel upper bound on the sampling error of CLD-based generative models in the Wasserstein metric. This additional hyperparameter influences the smoothness of sample paths, and our discretization error analysis provides practical guidance for its tuning, leading to improved sampling performance.

1 Introduction

Recent surge in machine learning and artificial intelligence has driven substantial progress in generative modeling, particularly with the development of Score-based Generative Models (SGMs). These models build on the foundational works in Denoising Diffusion Probabilistic Models (DDPMs) by Sohl-Dickstein et al. (2015); Song and Ermon (2019); Ho et al. (2020) and the advances in score-matching techniques introduced by Hyvärinen and Dayan (2005); Vincent (2011).

Score-based Generative Models (SGMs). SGMs are probabilistic models designed to create synthetic instances of a target distribution when only a genuine sample (*e.g.*, a dataset of real-life images) is accessible. First, the forward process involves progressively perturbing the training distribution by adding noise to the data until its distribution approximately reaches an easy-to-sample distribution π_∞. Then, the backward process involves learning to reverse this noising dynamics by sequentially removing the noise. SGMs have quickly gained recognition for their ability to generate high-quality synthetic data. Their applications span diverse areas, including computer vision (Li et al., 2022; Lugmayr et al., 2022), natural language processing (Gong et al., 2023), and other domains where realistic data generation is crucial. This growing body of work has been comprehensively surveyed by Yang et al. (2023), highlighting the versatility and potential of diffusion models. In

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

addition, SGMs provide a particularly interesting class of prior distributions to solve Bayesian inverse problems. Although they lack an explicit and tractable probability density function, a very active research area focuses on combining Monte Carlo guidance and SGMs to solve posterior sampling problems, Wu et al. (2023); Moufad et al. (2025); Victorino Cardoso et al. (2024).

Critically-damped Langevin Diffusion (CLD). In Dockhorn et al. (2022), the authors proposed Critically-damped Langevin Diffusion as a second-order extension of conventional diffusion models. By introducing velocity variables alongside the usual state variables —much like in Hamiltonian Monte Carlo— CLD accelerates exploration of high-dimensional spaces and often yields better sample quality in practice. Although empirical work demonstrates the benefit of CLD over standard score-based models (Dockhorn et al., 2022), its theoretical underpinnings remain incomplete. Existing convergence guarantees are only expressed in terms of Kullback–Leibler divergence (Conforti et al., 2025; Chen et al., 2023) and fail to capture any computational advantage for kinetic dynamics, leaving a gap between observed performance and formal analysis.

Contributions. We first discuss the challenges of establishing Wasserstein convergence under the standard assumptions used for Variance-Preserving (VP) or Variance-Exploding (VE) SGMs, where the forward process is elliptic (Gao et al., 2025; Strasman et al., 2025; Gentiloni-Silveri and Ocello, 2025; Bruno et al., 2025). We then provide, to the best of our knowledge, the first upper bound for CLD in the Wasserstein metric through coupling techniques under weaker assumptions, achieving convergence rates comparable to those of other SGMs. Crucially, this result is not implied by previous Kullback–Leibler divergence bounds (Conforti et al., 2025; Chen et al., 2023), and our proof technique differs significantly from existing Wasserstein analyses of diffusion models.

However, it is possible to introduce a modified dynamics that includes an additional hyperparameter controlling the noise on the data coordinate of CLD, thereby restoring ellipticity and enabling an analysis closely aligned with that of VP and VE models, but formulated on an extended phase space with matrix-valued drifts and diffusions. This hyperparameter governs the smoothness of sample paths, allowing a detailed analysis of the generative error as a function of this smoothness parameter. Such analysis offers practical guidance for tuning this hyperparameter and potentially improves sampling performance compared to standard SGMs and CLD methods. The benefits of this additional parameterization are demonstrated numerically on challenging synthetic datasets.

2 Notation and Background

Notation. We use π to denote probability distributions and p to denote their corresponding densities with respect to the Lebesgue measure or another reference measure. The identity matrix of size d is written $\mathbf{I}_d$. For $x, y \in \mathbb{R}^d$, we denote by $\langle x, y\rangle$ the standard inner product of $\mathbb{R}^d$, by $\|\cdot\|$ the Euclidean norm for vectors and its induced operator norm for matrices. Let $\|\cdot\|_F$ be the Frobenius norm defined for $A \in \mathbb{R}^{d\times d}$ as $\|A\|_F := \sqrt{\mathrm{Tr}(A^\top A)}$. For symmetric matrices $A, B \in \mathbb{R}^{d\times d}$, we write $A \preccurlyeq B$ to mean that $B - A$ is positive semidefinite. We denote the time derivative of a function by $\dot{f}(t) := \frac{\mathrm{d}}{\mathrm{d}t} f(t)$. We use the symbol $\otimes$ either for the Kronecker product when applied to matrices and for the product of probability measures when applied to distributions. The intended meaning will be clear from context. For any matrix $A \in \mathbb{R}^{d\times d}$ we denote its largest eigenvalue (resp. singular value) by $\lambda_{\max}(A)$ (resp. $\sigma_{\max}(A)$) and smallest eigenvalue by $\lambda_{\min}(A)$ (resp. $\sigma_{\min}(A)$). For random vectors $X, Y \in \mathbb{R}^d$, define $\|X\|_{L_2} := \left(\mathbb{E}\left[\|X\|^2\right]\right)^{1/2}$ and we write $X \perp Y$ to mean that X is independent of Y. The notation $\mathcal{L}(X)$ denotes the law (distribution) of a random vector X. For $a, b \in \mathbb{R}$, we write $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$.

Score-based Generative Models. SGMs employ a Gaussian Markovian diffusion process that smoothly transports the target data distribution $\pi_{\mathrm{data}} \in \mathcal{P}(\mathbb{R}^d)$ towards an easy-to-sample Gaussian distribution $p_\infty \in \mathcal{P}(\mathbb{R}^d)$. This process, known as forward diffusion, is the solution to the following stochastic differential equation (SDE) on a fixed time horizon $t \in [0, T]$,

$$\mathrm{d}\overrightarrow{X}_t = -\alpha\beta(t)\overrightarrow{X}_t\mathrm{d}t + \sqrt{2\beta(t)}\mathrm{d}B_t, \quad \overrightarrow{X}_0 \sim \pi_{\text{data}}, \tag{1}$$

with $(B_t)_{t\in[0,T]}$ a d-dimensional Brownian motion and $\beta(t) : [0, T] \to \mathbb{R}_+$. In particular, when $\alpha = 0$ and $\beta(t)$ is of the form $\beta^{\mathrm{VE}}(t)\dot{\beta}^{\mathrm{VE}}(t)$ the process is known as Variance Exploding (Song

and Ermon, 2019) and when $\alpha = 1$, the process is known as Variance Preserving (Sohl-Dickstein et al., 2015; Ho et al., 2020). This transformation can be reversed (Anderson, 1982; Haussmann and Pardoux, 1986; Cattiaux et al., 2023) and is also governed by an SDE, known as the backward process

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

where p_t is the time marginal p.d.f. of the forward process for $0 \leq t \leq T$. As a consequence, $\overleftarrow{X}_T$ has the same distribution as π_{data}. In practice, however, one cannot draw exact i.i.d. samples from this continuous-time process, and implementations of SGMs rely on three key approximations.

- *Mixing error*. The distribution of $\overrightarrow{X}_T$ is not analytically available in most cases, $\overleftarrow{X}_0$ is initialized at a known distribution π_∞, close to p_T.
- *Discretization error.* In most cases, the backward dynamic is non-linear, the backward process is discretized to sample from $\overleftarrow{X}_T$, which introduces an error due to evaluating the (time-continuous) score function only at discrete time steps.
- *Approximation error.* The score function depends on the unknown data distribution and thus cannot be computed in closed form. To approximate it, we use a neural network architecture $s_\theta : [0,T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ parameterized by $\theta \in \Theta$, and trained, for example, via Denoising Score Matching (see, *e.g.*, Vincent, 2011):

$$\mathcal{L}_{\mathrm{DSM}}(\theta) = \mathbb{E}\left[\lambda(t)\left\|s_\theta\left(\tau, \overrightarrow{X}_\tau\right) - \nabla \log p_\tau\left(\overrightarrow{X}_\tau | \overrightarrow{X}_0\right)\right\|^2\right], \tag{3}$$

  where τ is uniformly distributed on $[0,T]$, τ is independent of $\overrightarrow{X}_0$, $\overrightarrow{X}_\tau \sim p_\tau(\cdot|\overrightarrow{X}_0)$ and $\lambda : [0,T] \to \mathbb{R}_{>0}$ is a positive weighting function.

Theoretical studies of SGMs focus on those sources of errors to derive results for the total variation distance (De Bortoli et al., 2021), the Kullback–Leibler divergence (Conforti et al., 2025; De Bortoli et al., 2021; Chen et al., 2023; Benton et al., 2024) or the Wasserstein-2 distance (Lee et al., 2022, 2023; Bruno et al., 2025; Gao et al., 2025; Strasman et al., 2025; Gentiloni-Silveri and Ocello, 2025).

Kinetic Ornstein–Uhlenbeck. Inspired by Hamiltonian mechanics, kinetic SGMs operate in an extended position-velocity phase space, defined as $\overrightarrow{\mathbf{U}}_t = (\overrightarrow{X}_t, \overrightarrow{V}_t)^\top \in \mathbb{R}^{2d}$ which satisfies the following stochastic differential equation

$$\mathrm{d}\overrightarrow{\mathbf{U}}_t = A\overrightarrow{\mathbf{U}}_t\mathrm{d}t + \Sigma\mathrm{d}B_t\,, \qquad \overrightarrow{\mathbf{U}}_0 \sim \pi_{\mathrm{data}} \otimes \pi_v\,, \tag{4}$$

where $\pi_v \sim \mathcal{N}(0, v^2\mathbf{I}_d)$, $(B_t)_{t\in[0,T]}$ denotes a $2d$-dimensional standard Brownian motion,

$$A = \begin{pmatrix} 0 & a^2 \\ -1 & -2a \end{pmatrix} \otimes \mathbf{I}_d\,, \quad \text{and} \quad \Sigma = \begin{pmatrix} 0 & 0 \\ 0 & \sigma \end{pmatrix} \otimes \mathbf{I}_d\,. \tag{5}$$

Similar to (1), this process is Gaussian conditional on the distribution at time 0 (see Proposition A.2). The associated linear system corresponds to the stochastic analogue of a damped harmonic oscillator in the critically damped regime, with $a = 1/\sqrt{M}$ and $\sigma = 2/\sqrt{a}$, following the parameterization of Dockhorn et al. (2022). Note that (4) can also be expressed using a time-change or noise-schedule function $\beta : [0,T] \to \mathbb{R}_+$ (see Section E.2) . This will not play a key role in our theoretical analysis but is an important feature of practical numerical implementation.

Applying time-reversal results for diffusion processes (see, $e.g$, Haussmann and Pardoux, 1986; Cattiaux et al., 2023), the backward process $(\overleftarrow{\mathbf{U}}_t)_{t\geq 0}$ is solution to the following SDE:

$$\mathrm{d}\overleftarrow{\mathbf{U}}_t = -A\overleftarrow{\mathbf{U}}_t\mathrm{d}t + \Sigma^2\nabla \log p_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)\mathrm{d}t + \Sigma\mathrm{d}B_t\,, \tag{6}$$

with initial condition $\overleftarrow{\mathbf{U}}_0 \sim p_T$, where $p_t : \mathbb{R}^{2d} \to \mathbb{R}_+$ is the probability density function of $\overrightarrow{\mathbf{U}}_t$.

CLD-based SGMs. To sample from $\overleftarrow{\mathbf{U}}_t$ (and, in particular, from $\overleftarrow{X}_T \sim \pi_{\mathrm{data}}$), one must rely on the three SGM approximations discussed earlier. The *mixing error* is analogous to that of standard SGMs and leverages the ergodicity of the forward process—converging to a known Gaussian

distribution—to initialize the backward process. The *discretization* of the nonlinear backward SDE can be performed using classical numerical integrators commonly employed in SGMs, such as Euler–Maruyama (Song et al., 2021) or exponential integrators (Conforti et al., 2025). Additionally, due to the Hamiltonian structure of the kinetic process, symplectic integrators (Neal, 2011) may also be appropriate (Dockhorn et al., 2022). Finally, the *score approximation* can be implemented by applying Denoising Score Matching–similar to (3)–on the extended phase space $\overrightarrow{\mathbf{U}}_t = (\overrightarrow{X}_t, \overrightarrow{V}_t)^\top$, that is, using the conditional score function $\nabla \log p_t(\overrightarrow{\mathbf{U}}_t | \overrightarrow{\mathbf{U}}_0)$. However, since the distribution of $\overrightarrow{V}_0$ is known and Gaussian, it can be analytically marginalized, yielding the following objective function known as Hybrid Score Matching:

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

where τ is uniformly distributed on $[0,T]$, $\tau \perp \overrightarrow{X}_0$, $\overrightarrow{\mathbf{U}}_\tau \sim p_\tau(\cdot|\overrightarrow{X}_0)$ and $\lambda : [0,T] \to \mathbb{R}_{>0}$ is a positive weighting function. Empirically, Hybrid Score Matching tends to yield more stable training dynamics by reducing the variance of the training objective (Dockhorn et al., 2022).

3 Wasserstein Convergence of CLDs

In this section, we analyze the convergence of CLDs with respect to the 2-Wasserstein distance under the Euler–Maruyama discretization scheme. We first discuss the motivation for this analysis before introducing the setting, assumptions, and main results.

3.1 Motivation

While convergence results have been established in terms of the Kullback–Leibler divergence (Conforti et al., 2025; Chen et al., 2023), no analogous results currently exist for the Wasserstein-2 metric. Proving convergence in $\mathcal{W}_2$ requires establishing a contraction property of the backward dynamics in this metric—a challenging task for hypo-coercive SDEs (Villani, 2009; Eberle et al., 2019; Monmarché, 2023). The main difficulty arises from the degeneracy of the diffusion term, since the Brownian motion in CLDs acts only on the velocity component. To illustrate this point, consider the following example.

Introduce the change of variables $\overrightarrow{Y}_t = \overrightarrow{X}_t + a\overrightarrow{V}_t$, under which one component of the system evolves as an Ornstein–Uhlenbeck process. Writing $\overrightarrow{Z}_t = (\overrightarrow{X}_t, \overrightarrow{Y}_t)^\top$, the forward SDE in (4) can be rewritten as

$$\mathrm{d}\overrightarrow{Z}_t = a\begin{pmatrix}-1 & 1\\ 0 & -1\end{pmatrix}\overrightarrow{Z}_t \mathrm{d}t + \begin{pmatrix}0 & 0\\ 0 & \sigma\end{pmatrix}\mathrm{d}B_t\,.$$

Notably, the transformed process $(\overrightarrow{Y}_t)_{t\in[0,T]}$ corresponds to an Ornstein–Uhlenbeck process. By the time-reversal property, the corresponding backward process satisfies

$$\overleftarrow{Y}_t = \overleftarrow{X}_t + a\overleftarrow{V}_t\,,$$

which leads to the following backward SDE:

$$\mathrm{d}\overleftarrow{Z}_t = a\begin{pmatrix}1 & -1\\ 0 & 1\end{pmatrix}\overleftarrow{Z}_t \mathrm{d}t + \sigma^2\begin{pmatrix}0\\ \nabla_y \log p_{T-t}\left(\overleftarrow{Z}_t\right)\end{pmatrix}\mathrm{d}t + \begin{pmatrix}0 & 0\\ 0 & \sigma\end{pmatrix}\mathrm{d}B_t\,, \tag{7}$$

where p_t denotes the probability density function of $\overrightarrow{Z}_t$. A standard approach to establishing contraction consists in studying the difference process associated with the dynamics in (7), starting from two deterministic initial conditions $(x_0, y_0), (x_0', y_0') \in \mathbb{R}^{2d}$ and denoting by $(X_t, Y_t)_{t\in[0,T]}$ and $(X_t', Y_t')_{t\in[0,T]}$ the corresponding solutions. Under a synchronous coupling–*i.e.* using the same Brownian motion to drive the evolution of both processes–the difference process becomes a deterministic ODE, whose stability properties determine the contraction properties of the system. In particular, using the mean value theorem applied to the gradients of the log-density, the following holds for $t \in [0;T)$:

$$\mathrm{d}\begin{pmatrix}X_t - X_t'\\ Y_t - Y_t'\end{pmatrix} = \begin{pmatrix}a + \sigma^2 G_t & -a\\ 0 & a + \sigma^2 H_t\end{pmatrix}\begin{pmatrix}X_t - X_t'\\ Y_t - Y_t'\end{pmatrix}\mathrm{d}t\,. \tag{8}$$

where

$$H_t = \int_0^1 \nabla_y^2 \log p_{T-t}\big(X'_t + \gamma(X_t - X'_t),\ Y'_t + \gamma(Y_t - Y'_t)\big)\,\mathrm{d}\gamma\,,$$
$$G_t = \int_0^1 \nabla_y \nabla_x^\top \log p_{T-t}\big(X'_t + \gamma(X_t - X'_t),\ Y'_t + \gamma(Y_t - Y'_t)\big)\,\mathrm{d}\gamma\,.$$

To ensure contraction of the system, all eigenvalues of the matrix in (8) must be negative. However, the main difficulty lies in controlling the term G_t, which involves the mixed second-order derivative $\nabla_y \nabla_x^\top \log p_{T-t}$. For contraction to occur, this term must also be sufficiently negative. This is a strong and challenging requirement, as it demands a form of joint concavity of cross-derivatives, which is not generally ensured even when p_{T-t} is strongly log-concave in each variable separately.

3.2 Settings: Dynamics and Backward Discretization

Position-noise regularization in the extended phase space. As detailed in Dalalyan and Riou-Durand (2020), kinetic Langevin-based samplers depend on the mixing rate and on the regularity of the underlying dynamics. To better exploit the extended phase space, we introduce a modified dynamics that adds a small noise term on the position coordinate $\varepsilon \geq 0$. Crucially, when ε is strictly positive, this modification restores ellipticity of the forward and backward processes, which facilitates greatly the theoretical analysis. This hyperparameter controls the smoothness of the sample paths and the analysis of the discretization error allows a practical tuning to improve sampling performance in comparison with standard SGM models and kinetic-based diffusion samplers. The diffusion coefficient of the forward SDE is then given by

$$\Sigma_\varepsilon := \begin{pmatrix} \varepsilon & 0 \\ 0 & \sigma \end{pmatrix} \otimes \mathbf{I}_d\,,$$

giving a process $(\overrightarrow{\mathbf{U}}_t)_{t\in[0,T]} \in \mathbb{R}^{2d}$ which satisfies the following SDE

$$\mathrm{d}\overrightarrow{\mathbf{U}}_t = A\overrightarrow{\mathbf{U}}_t \mathrm{d}t + \Sigma_\varepsilon \mathrm{d}B_t\,, \qquad \overrightarrow{\mathbf{U}}_0 \sim \pi_{\mathrm{data}} \otimes \pi_v \tag{9}$$

with $\varepsilon \geq 0$. Note that the case $\varepsilon = 0$ recovers the classical CLD framework. In the following, we write

$$\mathsf{s}_t(u) = \nabla \log p_t(u)\,, \quad \text{for } t \geq 0,\ u \in \mathbb{R}^{2d}\,. \tag{10}$$

Modified score function. Following Conforti et al. (2025), we adopt a modified score formulation based on the rescaled density $\tilde{p}_t := p_t/p_\infty$, where p_∞ is the density of the stationary distribution associated with (4). This perspective, also emphasized in Cattiaux et al. (2023); Conforti and Léonard (2022); Strasman et al. (2025); Conforti et al. (2025); Gentiloni-Silveri and Ocello (2025); Pham et al. (2025), reveals deep connections with stochastic control theory. In particular, the modified score satisfies a Hamilton–Jacobi–Bellman (HJB) equation, which we highlight and exploit in the sequel. With this notation, the backward process $\overleftarrow{\mathbf{U}}$ can be written equivalently as

$$\mathrm{d}\overleftarrow{\mathbf{U}}_t = \tilde{A}_\epsilon \overleftarrow{\mathbf{U}}_t \mathrm{d}t + \Sigma_\varepsilon^2 \nabla \log \tilde{p}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)\mathrm{d}t + \Sigma_\varepsilon \mathrm{d}B_t\,, \tag{11}$$

with $\tilde{A}_\epsilon = -A - \Sigma_\epsilon^2 \Sigma_\infty^{-1}$. In the following, we write $\tilde{\mathsf{s}}_t(u) := \nabla \log \tilde{p}_t(u)$, for $t \geq 0,\ u \in \mathbb{R}^{2d}$.

Backward process discretization. Let $N \in \mathbb{N}$ denote the number of discretization steps, so that $0 = t_0 < t_1 < \ldots < t_N = T$. To analyze the convergence of the discretized backward process, we introduce the continuous-time interpolation $(\bar{\mathbf{U}}_t)_{t\in[0,T]}$ of the Euler scheme for the time-reversed process $(\overleftarrow{\mathbf{U}}_t)_{t\in[0,T]}$. This is defined as the Itô process such that, for $t \in [t_k, t_{k+1}]$,

$$\bar{\mathbf{U}}_t = \bar{\mathbf{U}}_{t_k} + \left(\tilde{A}_\epsilon \bar{\mathbf{U}}_{t_k} + \Sigma_\varepsilon^2 \tilde{\mathsf{s}}_{T-t_k}\big(\bar{\mathbf{U}}_{t_k}\big)\right)(t - t_k) + \Sigma_\varepsilon\big(B_t - B_{t_k}\big)\,, \tag{12}$$

where the process is initialized at p_T (i.e., $\bar{\mathbf{U}}_0 \sim p_T$). When initialized at π_∞, we denote by $(\bar{\mathbf{U}}_t^\infty)_{t\in[0,T]}$ the corresponding Itô process. For simplicity, the discretization is performed on a uniform grid, with step size $h = T/N$, so that $t_{k+1} - t_k = h$ for all k.

Generative model. The *generative model* is defined as the continuous-time interpolation of the discretized backward process, in which the true (unknown) modified score function is replaced by its parametric approximation $\tilde{s}_\theta : [0,T] \times \mathbb{R}^{2d} \mapsto \mathbb{R}^d$. The resulting process, denoted by $(\bar{\mathbf{U}}_t^\theta)_{t\in[0,T]}$, satisfies for $t \in [t_k, t_{k+1}]$

$$\bar{\mathbf{U}}_t^\theta = \bar{\mathbf{U}}_{t_k}^\theta + \left(\tilde{A}_\epsilon \bar{\mathbf{U}}_{t_k}^\theta + \Sigma_\varepsilon^2 \tilde{s}_\theta\big(t_k, \bar{\mathbf{U}}_{t_k}^\theta\big)\right)(t - t_k) + \Sigma_\varepsilon\big(B_t - B_{t_k}\big)\,, \tag{13}$$

with initialization $\bar{\mathbf{U}}_0^\theta \sim \pi_\infty$. Learning the modified score function $\nabla \log \tilde{p}_t$ is theoretically equivalent to learning the standard score function $\nabla \log p_t$, since the two functions differ only by a known linear term. As a consequence, the modified score approximation can be written, for any $t \geq 0$ and any $u \in \mathbb{R}^{2d}$, as

$$\tilde{s}_\theta(t,u) := s_\theta(t,u) + \Sigma_\infty^{-1} u\,.$$

Ultimately, the objective is to control the $\mathcal{W}_2$–distance between $\mathcal{L}(\bar{X}_T^\theta)$ the generated data marginal distribution at time T and π_{data} the true data distribution (recall that $\bar{\mathbf{U}}_T^\theta = (\bar{X}_T^\theta, \bar{V}_T^\theta)^\top$).

3.3 Assumptions

Regularity assumptions.

H1 The data distribution π_{data} is absolutely continuous w.r.t. the Lebesgue measure, with density p_{data} and the relative Fisher information between $\pi_0 = \pi_{\mathrm{data}} \otimes \pi_v$ (*i.e.* the initialization of the stochastic process defined in (4)) and π_∞ is finite, *i.e.*

$$\mathcal{I}(\pi_0|\pi_\infty) := \int \left\| \nabla \log\left(\frac{\mathrm{d}\pi_0}{\mathrm{d}\pi_\infty}(u)\right)\right\|^2 \pi_0(\mathrm{d}u) < \infty\,.$$

Assumption H1, particularly the requirement of finite Fisher information, is standard in most works establishing convergence bounds for SGMs. This condition is either explicitly assumed or implied by stronger regularity assumptions used in the literature (Conforti et al., 2025; Strasman et al., 2025).

H2 (i) The data distribution is of the form $p_{\mathrm{data}}(x) \propto \exp\left(-(V(x) + H(x))\right)$ and satisfies:
- There exists $L > 0$ such that $|H(x) - H(y)| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$.
- There exists $\alpha > 0$ such that $\alpha \mathbf{I}_d \preceq \nabla^2 V(x)$ for all $x \in \mathbb{R}^d$.

(ii) $(-\log p_{\mathrm{data}})$ is L_0-one-sided Lipschitz, *i.e.*, for all $x, y \in \mathbb{R}^d$,

$$-\left(\nabla \log p_{\mathrm{data}}(x) - \nabla \log p_{\mathrm{data}}(y)\right)^\top (x - y) \leq L_0 \|x - y\|^2\,. \tag{14}$$

The first point of Assumption H2 models the data distribution as a strongly log-concave component V perturbed by a term H, similar to the settings considered in Brigati and Pedrotti (2025); Stéphanovitch (2025). Intuitively, this assumption allows the distribution to deviate from strong log-concavity via the perturbation H, while still maintaining sufficient regularity for the analysis. When $H = 0$, the distribution reduces to the strongly log-concave case, which is commonly used to establish contraction in the Wasserstein metric (Bruno et al., 2025; Gao et al., 2025; Strasman et al., 2025). The second point of Assumption H2 assumes a one-sided Lipschitz condition, which is weaker than requiring full Lipschitz continuity of the score function (Gentiloni-Silveri and Ocello, 2025). Notably, H2 implies the Lipschitz continuity of the score function. This means that for all $t \in (0,T]$, there exists $L_t > 0$ such that for all $u, \bar{u} \in \mathbb{R}^{2d}$,

$$\|\mathsf{s}_t(u) - \mathsf{s}_t(\bar{u})\| \leq L_t \|u - \bar{u}\|\ . \tag{15}$$

This condition can be verified under standard assumptions. In particular, if $\nabla \log p_{\mathrm{data}}$ is Lipschitz, the assumption holds. Since π_{data} and π_v are independent and π_v is often Gaussian, it suffices to assume that p_{data} is log-smooth, a common condition in the analysis of SGMs to ensure convergence (Gao et al., 2025; Chen et al., 2023).

Furthermore, Assumption H2 ensures that π_{data} has sub-Gaussian tails (Lemma D.1). Consequently, all its polynomial moments are finite. In particular, π_{data} admits a finite second moment, a standard condition—either explicit or implied by stronger regularity assumptions—in convergence analyses of SGMs. Importantly, Assumption H2, together with the polynomial growth condition $\|\nabla V(x)\| \leq C(1 + \|x\|^m)$ for all $x \in \mathbb{R}^d$, with some $C > 0$ and $m \in \mathbb{N}$, implies Assumption H1 (Lemma D.2).

These assumptions are satisfied by standard distributions such as Gaussian and mixtures of Gaussians. They are strictly weaker than the conditions typically required in the literature to establish Wasserstein convergence guarantees—such as strong log-concavity combined with the Lipschitz continuity of the score function (Gao et al., 2025; Strasman et al., 2025)—which hold only for non-degenerate Gaussian distributions and therefore exclude many practically relevant settings, even though they remain common in the literature.

Score approximation.

H3 There exists $M \geq 0$ such that,

$$\sup_{k\in\{0,..,N-1\}} \left\|\tilde{\mathbf{s}}_{T-t_k}\left(\bar{\mathbf{U}}^\theta_{t_k}\right) - \tilde{s}_\theta\left(T-t_k, \bar{\mathbf{U}}^\theta_{t_k}\right)\right\|_{L_2} \leq M\,.$$

Assumption H3 is standard in the literature (De Bortoli et al., 2021; Conforti and Léonard, 2022; Gao et al., 2025; Bruno et al., 2025; Strasman et al., 2025; Gentiloni-Silveri and Ocello, 2025; Cordero-Encinar et al., 2025) as essentially all convergence proofs for diffusion-based score models require that the neural network has learned the score within some uniform error. This condition quantifies the ability of the neural network architecture to approximate the true score function and serves to control the score approximation error.

3.4 Main Results

We establish here the Wasserstein-2 convergence of CLD-based SGMs under these weak assumptions. A key step is to show that, under Assumption H2, the scaled score function $\Sigma_\varepsilon^2\nabla\log p_t$ (resp. $\Sigma_\varepsilon^2\nabla\log\tilde{p}_t$) is L_t-Lipschitz (resp. $\tilde{L}_t$-Lipschitz), for $t>0$ (Proposition B.1). This, in particular, yields an exponential decay of the operator norm $\|\Sigma_\varepsilon^2\nabla^2\log\tilde{p}_t\|$ as $t\to\infty$. The following theorem provides, to the best of our knowledge, the first convergence rates in Wasserstein distance for CLD-based approaches and aligns with recent developments in the literature of Variance-Preserving and Variance-Exploding SGMs.

Theorem 3.1. *Assume that Assumptions H1- H3 hold. Then, there exist $c_1, c_2 > 0$ such that, for all $h > 0$,*

$$\mathcal{W}_2\left(\pi_{\mathrm{data}}, \mathcal{L}\left(\bar{X}^\theta_T\right)\right) \leq c_1\mathrm{e}^{-c_2T}\mathcal{W}_2\left(\pi_{\mathrm{data}}\otimes\pi_v, \pi_\infty\right) + c_1\sigma^2 M + c_1\sqrt{h}\,.$$

Proof. Let $P_X : \mathbb{R}^{2d}\to\mathbb{R}^d$ denote the projection $P_X(x,v)=x$. Using that P_X is 1–Lipschitz for the Euclidean norm, yields,

$$\mathcal{W}_2\big(\pi_{\mathrm{data}},\, \mathcal{L}\big(\bar{X}^\theta_T\big)\big) \leq \mathcal{W}_2\big(\pi_{\mathrm{data}}\otimes\pi_v,\, \mathcal{L}\big(\bar{\mathbf{U}}^\theta_T\big)\big)\,. \tag{16}$$

The right-hand side of (16) is then bounded by decomposing the total generation error, using the triangle inequality, into the three sources of error for SGMs discussed in Section 2:

$$\begin{aligned}\mathcal{W}_2\left(\pi_{\mathrm{data}}\otimes\pi_v, \mathcal{L}\left(\bar{\mathbf{U}}^\theta_T\right)\right) \leq{}& \mathcal{W}_2\left(\mathcal{L}\left(\overleftarrow{\bar{\mathbf{U}}}_T\right), \mathcal{L}\left(\bar{\mathbf{U}}_T\right)\right) + \mathcal{W}_2\left(\mathcal{L}\left(\bar{\mathbf{U}}^\infty_T\right), \mathcal{L}\left(\bar{\mathbf{U}}^\theta_T\right)\right)\\ &+ \mathcal{W}_2\left(\mathcal{L}\left(\bar{\mathbf{U}}_T\right), \mathcal{L}\left(\bar{\mathbf{U}}^\infty_T\right)\right)\,,\end{aligned}$$

where $\bar{\mathbf{U}}_T$ and $\bar{\mathbf{U}}^\infty_T$ are defined in Equation (12) and $\bar{\mathbf{U}}^\theta_T$ in Equation (13). The first term (discretization error) is controlled by Lemma B.2, which ensures that there exists $c_1>0$ such that, for all $h>0$,

$$\mathcal{W}_2(\mathcal{L}(\overleftarrow{\bar{\mathbf{U}}}_T), \mathcal{L}(\bar{\mathbf{U}}_T)) \leq c_1\sqrt{h}\,.$$

The second term (score approximation error) is bounded by Lemma B.3,

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{\mathbf{U}}^\infty_T\right), \mathcal{L}\left(\bar{\mathbf{U}}^\theta_T\right)\right) \leq c_1\sigma^2 M\,.$$

Finally, the third term (mixing error) is controlled by Lemma B.4, which guarantees the existence of $c_2>0$ such that,

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{\mathbf{U}}_T\right), \mathcal{L}\left(\bar{\mathbf{U}}^\infty_T\right)\right) \leq c_1\mathrm{e}^{-c_2T}\mathcal{W}_2\left(\pi_{\mathrm{data}}\otimes\pi_v, \pi_\infty\right)\,.$$

Combining these three bounds together with (16) concludes the proof. □

Theorem 3.1 establishes convergence rates in the Wasserstein distance for CLD-based approaches for all $\epsilon \geq 0$, recovering the vanilla CLD when $\epsilon = 0$. In this case, our result aligns with the KL convergence analyses of kinetic Langevin dynamics by Chen et al. (2023) and Conforti et al. (2025) for the specific choice $a = 1$ and $\sigma = 2$. It is worth emphasizing that, under our weaker assumptions, no equivalence holds between KL and Wasserstein convergence, so our results are not implied by existing KL-based analyses. Beyond this theoretical bound, our analysis indicates that smaller values of v yield better log-concavity constants; however, v is typically chosen to be small but not too small, to avoid an explosion in the Lipschitz constant. This remark is consistent with the empirical evidence brought forward by Dockhorn et al. (2022), which suggests that small values of v may improve training stability and sampling performance.

3.5 Strongly Log-Concave Case

This subsection focuses on the elliptic case, *i.e.*, when $\varepsilon > 0$. In this setting, the forward process associated with CLD becomes a *multidimensional Ornstein–Uhlenbeck process* with matrix-valued drift and diffusion coefficients. The presence of the additional noise term on the position coordinate restores ellipticity, which allows us to extend classical convergence analyses developed for VP and VE diffusions to this kinetic framework.

Crucially, in the strongly log-concave case, i.e., when $H = 0$ in Assumption H2, the upper bound can be expressed with more explicit constants that depend on the regularity of the data. Moreover, in this case, the one-sided Lipschitz condition becomes equivalent to the Lipschitz continuity of the score function. Assumption H2 then reduces to the following assumption.

H2′ The data distribution is absolutely continuous w.r.t. the Lebesgue measure, is of the form $p_{\mathrm{data}}(x) \propto \mathrm{e}^{-V(x)}$ and is α_0–strongly log-concave and L_0-log-smooth, *i.e.*, there exists $\alpha_0 > 0$ and $L_0 > 0$ such that,

$$\alpha_0 \mathbf{I}_d \preceq \nabla^2 V(x) \preceq L_0 \mathbf{I}_d, \qquad \text{for all } x \in \mathbb{R}^d \,.$$

Under this assumption, the forward flow preserves both strong log-concavity and smoothness. Indeed, Propositions C.1 and C.2 guarantee that p_t remains α_t–log-concave and L_t–log-smooth for all $t \in [0, T]$, with α_t and L_t explicitly defined as functions of α_0 and L_0 in the respective propositions. Such regularity properties are fundamental for proving exponential contraction in the Wasserstein metric, and are consistent with the analysis of classical (VP) diffusion models (Bruno et al., 2025; Gao et al., 2025; Strasman et al., 2025). In contrast, Chen et al. (2023) obtain Wasserstein convergence guarantees without requiring strong log-concavity, instead relying on the compactness of the domain and (15), a setting where convergence in KL divergence is effectively equivalent. The following theorem presents Wasserstein convergence results under assumptions for which no such equivalence with the KL divergence holds. In particular, our result is not implied by existing analyses based on KL convergence.

Theorem 3.2. *Assume that H2′ and H3 hold, and let $\varepsilon > 0$. If the step size h satisfies*

$$0 < h < \frac{2 \min_k \alpha_{t_k} \left(\sigma^2 \wedge \varepsilon^2\right) - (\sigma - \varepsilon)^2 \max_k L_{t_k} - (a+1)^2}{\|A\|^2 + (\varepsilon^4 + \sigma^4) \max_k L_{t_k}^2 + 2\left(\sigma^2 \vee \varepsilon^2\right) \|A\| \max_k L_{t_k}} \,, \tag{17}$$

then,

$$\mathcal{W}_2\left(\pi_{\mathrm{data}}, \mathcal{L}\left(\bar{X}_T^\theta\right)\right) \lesssim K_T \mathrm{e}^{-aT} \mathcal{W}_2\left(\pi_{\mathrm{data}} \otimes \pi_v, \pi_\infty\right) + \sigma^2 M + \sqrt{h}\, C_a(\varepsilon) \,.$$

where $K_T = (1 + \max\{a+1; a(a+1)\}T)$ and

$$C_a(\varepsilon) = \Big(2\|A\|^4 B_\varepsilon + 4d(a^2\sigma^2 + \varepsilon)^2 \Lambda_\varepsilon^*(T)\Big) h + 4d\Big(\|A\|^2 + \sigma^4 \sup_{t \in [0,T]} L_{T-t}^2\Big) \,,$$

with B_ε and $\Lambda_\varepsilon^*(T)$ as in Lemma C.3.

Proof. The error decomposition is the same as in Theorem 3.1. The full statement and proof for each error term is provided in Appendix C. □

This bound highlights the stabilizing role of the parameter $\varepsilon > 0$, which restores ellipticity in the dynamics. A key observation is that

$$\Sigma_\varepsilon \nabla^2 \log p_t \Sigma_\varepsilon \preccurlyeq -(\varepsilon^2 \wedge \sigma^2) \alpha_t \mathbf{I}_{2d} \,,$$

which can be negative only for positive values of ε. In this sense, increasing ε tends to enhance the contractive behavior of the dynamics, as also reflected by the admissible step-size condition (17). However, this effect is not purely beneficial: several terms in the discretization error scale with ε^2, illustrating that excessive noise injection may deteriorate the regularity of the process. Consequently, there is a trade-off in the choice of ε to balance these competing effects. This trade-off is numerically illustrated in Section 4.

Remark 3.3. Finite second order moment is also necessary in this approach and is deduced from H2′ (see, *e.g.*, Lemma B.1, Gentiloni-Silveri and Ocello, 2025). Regarding H3, it is implied that the score approximation is now made for the true score function, not the modified one.

4 Experiments

We illustrate the effect of the regularization parameter ε on the generation quality of CLDs on a simple yet challenging toy dataset. The regularization parameter ε is chosen to be in $\{0, 0.1, 0.25, 0.5, 1\}$. Notably, $\varepsilon = 0$ corresponds to the vanilla CLD setting. Our source code is publicly available here[1].

Evaluation metric. To assess the quality of the generated samples, directly computing the Wasserstein-2 distance is infeasible, as it requires solving a computationally expensive optimal transport problem. Instead, we approximate the $\mathcal{W}_2$-distance between the generated samples (with distribution $\hat{\pi}$) and the training samples (with distribution $\pi\mathrm{data}$) using the sliced Wasserstein distance (Flamary et al., 2021). It is defined as $SW_2^2(\pi_{\mathrm{data}}, \hat{\pi}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{U}(\mathbb{S}^{d-1})}[\mathcal{W}_2^2\left(\mathbf{u}_{\#}\pi_{\mathrm{data}}, \mathbf{u}_{\#}\hat{\pi}\right)]$ where $\mathcal{U}(\mathbb{S}^{d-1})$ denotes the uniform distribution over the unit sphere and $\mathbf{u}_{\#}$ is the push-forward operator associated with $\mathbf{u}$. The expectation is approximated using the standard Monte Carlo method with 2000 projections and π_{data} and $\hat{\pi}$ are replaced by their empirical distributions.

Dataset. We evaluate the generation quality on the Funnel distribution, which is characterized by a strong imbalance in variance across dimensions and was previously used in Thin et al. (2021). To further illustrate our results, we extend the evaluation to two additional challenging toy datasets (Appendix E.5): MG-25 (a 25-mode, 100-dimensional Gaussian mixture) and Diamonds (a 2-dimensional Gaussian mixture with a diamond-shaped geometry).

Hybrid Score Matching. Following the insights of Ho et al. (2020), the networks are trained to predict the noise (or rescaled noise) added during the forward process. When $\varepsilon = 0$, we use the positive weighting function proposed by Dockhorn et al. (2022) (see page 5, $\lambda(t) = \ell_t^{-2}$, in Dockhorn et al. (2022)). A similar reweighting, however, is not feasible for $\varepsilon \neq 0$ due to the matrix-valued nature of the objective function. Empirically, we observe that much of the training variance arises from the determinant computation involved in the 2×2 matrix inversions. To mitigate this, we set $\lambda(t) = \det(\Sigma_{0,t})^2$, which effectively stabilizes training. We parameterize the score network as $s_\theta(\overrightarrow{\mathbf{U}}_t, t) := -\Sigma_{0,t}^{-1}\alpha_\theta(\overrightarrow{\mathbf{U}}_t, t)$ so that the hybrid score matching objective for $\varepsilon > 0$ is given by

$$\mathcal{L}_{(\mathrm{HSM})^\varepsilon}(\theta) = \mathbb{E}\left[\det(\Sigma_{0,t})^2 \left\|\Sigma_{0,t}^{-1}\left(s_\theta\left(\tau, \overrightarrow{X}_\tau\right) - \Sigma_{0,t}^{1/2} G_{2d}\right)\right\|^2\right], \tag{18}$$

where G_{2d} denotes a $2d$-dimensional standard Gaussian noise.

Model, Training and Generation. All score networks share the same architecture: a fully connected neural network with three hidden layers of width 512 (see Figure 3). Training is performed using the Adam optimizer to minimize the hybrid score matching objective in (18), with a learning rate of 10^{-4} over 2000 epochs. The training set consists of 50 000 samples. For evaluation, we generate 50 000 samples using the Euler–Maruyama discretization scheme with $N = 1000$ steps and compare them against a test set of 50 000 samples. Both training and generation are independently repeated five times. The training (Algorithm 1) and sampling (Algorithm 2) procedures are provided in Appendix E.1.

Effect of the regularization parameter. Figure 1 illustrates the Wasserstein error for different values of the regularization parameter $\varepsilon \in \{0, 0.1, 0.25, 0.5, 1\}$ and drift coefficient $a \in \{0.1, 0.25, 0.5, 1, 2\}$. Across all values of a, introducing a small regularization parameter ε notably improves generation quality, even though the score network in the regularized case must predict a vector twice as long as in the non-regularized one. Moreover, regularization consistently reduces the

[1] https://github.com/SobihanSurendran/CLD

variance across runs. For smaller values of a, the error increases sharply when $\varepsilon = 0$ and also for large ε values. In contrast, for moderate values of a, the error becomes negligible, with $\varepsilon \in [0.1, 0.5]$ yielding slightly better performance than the other settings. It is worth noting that our experimental configuration closely follows that of Dockhorn et al. (2022), using $\sigma = \sqrt{2}$, $a = 2$, and in particular $\varepsilon = 0$. This observation justifies their choice of $a = 2$ for the vanilla CLD.

Effect of ε in controlled settings.

Varying ε modifies both the stationary distribution and the noise schedule—two factors known to strongly influence performance (Guo et al., 2023; Chen et al., 2023; Strasman et al., 2025)—it is important to control for these effects. To mitigate this confounding factor, one can fix the stationary distribution of the base case to $\mathcal{N}(0_{2d}, \mathbf{I}_{2d})$ and maintain comparable noise levels in the position and velocity spaces by setting $a(\varepsilon) = 1-\varepsilon^2/2$ and $\sigma(\varepsilon) = \sqrt{4+\varepsilon^2}$. This choice ensures that the stationary distribution remains close to $\mathcal{N}(0_{2d}, \mathbf{I}_{2d})$ for small values of ε. Although this adjustment becomes less accurate for larger ε, there is no practical limitation preventing the use of higher regularization values.

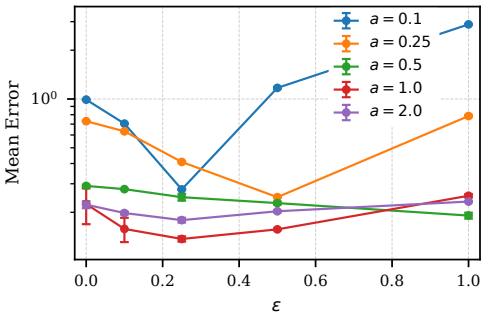

Figure 1: Mean $\mathcal{W}_2$ distance over 5 repetitions between the test set and generated samples on Funnel distribution in dimension 100. Error bars represent $\pm$ one standard deviation.

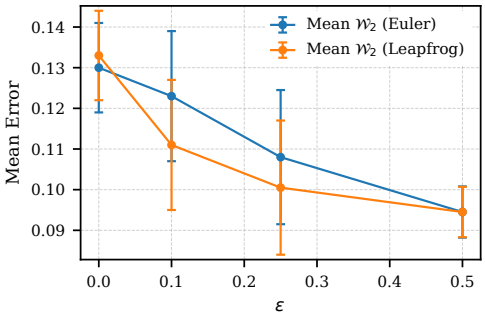

Figure 2: Mean $\mathcal{W}_2$ distance over 5 repetitions between the test set and generated samples on Funnel distribution with $a(\varepsilon) = 1 - \varepsilon^2/2$ and $\sigma(\varepsilon) = \sqrt{4+\varepsilon^2}$.

Figure 2 still shows an improvement in generation quality for small regularization parameters ε. To confirm that this effect is not tied to the discretization method, we reproduce the experiments using a Leapfrog integrator. As expected, the Leapfrog scheme outperforms Euler–Maruyama, yet the relative benefit of regularization persists. Finally, we emphasize that our objective is not to conduct an extensive numerical comparison of integrators or training strategies, but rather to highlight the potential of introducing controlled regularization within the CLD framework—a direction theoretically supported by Theorem 3.2.

5 Discussion

In this paper, we present the first theoretical analysis of the sampling error of CLDs in the Wasserstein metric under weaker assumptions than those previously used in the literature. Our results show that CLD-based samplers can achieve comparable convergence rates while effectively leveraging the structure of the extended space. We further analyze a generalized dynamic that extends classical CLDs by introducing a smoothness-controlling hyperparameter that regulates the noise on the data coordinate. This parameter provides more precise control over the regularity of sample paths and plays a central role in the discretization error analysis. Both theoretical and empirical results suggest that appropriately tuning this parameter leads to improved sampling quality and stability. Overall, our work offers both theoretical insights and practical guidance for CLD methods in generative modeling, particularly in scenarios where standard assumptions may not hold. Several promising directions remain for future research. Replacing the Euler discretization scheme with a higher-order method—such as the Leapfrog integrator, which is specifically designed for CLD-based dynamics—could further enhance sampling performance. Analyzing such schemes would likely yield sharper convergence rates consistent with the numerical results. Moreover, developing denoiser architectures specifically tailored to the extended space represents another promising avenue for applied research, potentially leading to tighter bounds on the approximation error.

Acknowledgements

The PhD of Sobihan Surendran was funded by the Paris Region PhD Fellowship Program of Région Ile-de-France. The work of Antonio Ocello was funded by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. We would also like to thank SCAI (Sorbonne Center for Artificial Intelligence) for providing the computing clusters.

References

Achleitner, F., Arnold, A., and Stürzer, D. (2015). Large-time behavior in non-symmetric fokker-planck equations. In *Rivista di Matematica della Università di Parme*, pages 1–68.

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

Benton, J., De Bortoli, V., Doucet, A., and Deligiannidis, G. (2024). Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations*.

Bouchut, F., James, F., and Mancini, S. (2005). Uniqueness and weak stability for multi-dimensional transport equations with one-sided lipschitz coefficient. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 4(1):1–25.

Brascamp, H. J. and Lieb, E. H. (1976). On extensions of the brunn–minkowski and prékopa–leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22:366–389.

Brigati, G. and Pedrotti, F. (2025). Heat flow, log-concavity, and lipschitz transport maps. *Electronic Communications in Probability*, 30:1–12.

Bruno, S., Zhang, Y., Lim, D.-Y., Akyildiz, Ö. D., and Sabanis, S. (2025). On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *Transactions on Machine Learning Research*.

Cattiaux, P., Conforti, G., Gentil, I., and Léonard, C. (2023). Time reversal of diffusion processes under a finite entropy condition. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 59(4):1844–1881.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. (2023). Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*.

Conforti, G., Durmus, A., and Silveri, M. G. (2025). KL convergence guarantees for score diffusion models under minimal data assumptions. *SIAM Journal on Mathematics of Data Science*, 7(1):86–109.

Conforti, G. and Léonard, C. (2022). Time reversal of markov processes with jumps under a finite entropy condition. *Stochastic Processes and their Applications*, 144:85–124.

Cordero-Encinar, P., Akyildiz, O. D., and Duncan, A. B. (2025). Non-asymptotic analysis of diffusion annealed langevin monte carlo for generative modelling. *arXiv preprint arXiv:2502.09306*.

Dalalyan, A. S. and Riou-Durand, L. (2020). On sampling from a log-concave density using kinetic langevin diffusions. *Bernoulli*, 26(3):1956–1988.

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

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

Eberle, A., Guillin, A., and Zimmer, R. (2019). Couplings and quantitative contraction rates for langevin dynamics. *The Annals of Probability*, 47(4):1982–2010.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Gao, X., Nguyen, H. M., and Zhu, L. (2025). Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, 26(43):1–54.

Gentiloni-Silveri, M. and Ocello, A. (2025). Beyond log-concavity and score regularity: Improved convergence bounds for score-based generative models in w2-distance. In *International Conference on Machine Learning*.

Gong, S., Li, M., Feng, J., Wu, Z., and Kong, L. (2023). Diffuseq: Sequence to sequence text generation with diffusion models. In *International Conference on Learning Representations*.

Guo, Q., Liu, S., Yu, Y., and Luo, P. (2023). Rethinking the noise schedule of diffusion-based generative models. *arXiv preprint arXiv:2309.12345*.

Haussmann, U. G. and Pardoux, E. (1986). Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, volume 33, pages 6840–6851.

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

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

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

Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., and Chen, Y. (2022). Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471.

Monmarché, P. (2023). Almost sure contraction for diffusions on $\mathbb{R}^d$. application to generalised langevin diffusions. *Stochastic Processes and their Applications*, 161:316–349.

Moufad, B., Janati, Y., Bedin, L., Durmus, A., Douc, R., Moulines, E., and Olsson, J. (2025). Variational diffusion posterior sampling with midpoint guidance. In *International Conference on Learning Representations*.

Neal, R. M. (2011). MCMC using hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, volume 54, chapter 5, pages 113–162. Chapman & Hall/CRC Press.

Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR.

Pham, L.-T.-N., Shariatian, D., Ocello, A., Conforti, G., and Durmus, A. (2025). Bit-level discrete diffusion with markov probabilistic models: An improved framework with sharp convergence bounds under minimal assumptions. In *International Conference on Machine Learning*.

Saumard, A. and Wellner, J. A. (2014). Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8(none):45 – 114.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pages 11918–11930.

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

Stéphanovitch, A. (2025). Regularity of the score function in generative models. *arXiv preprint arXiv:2506.19559*.

Strasman, S., Ocello, A., Boyer, C., Le Corff, S., and Lemaire, V. (2025). An analysis of the noise schedule for score-based generative models. *Transactions on Machine Learning Research*.

Thin, A., Janati El Idrissi, Y., Le Corff, S., Ollion, C., Moulines, E., Doucet, A., Durmus, A., and Robert, C. (2021). NEO: Non equilibrium sampling on the orbits of a deterministic transform. In *Advances in Neural Information Processing Systems*, volume 34, pages 17060–17071.

Victorino Cardoso, G., Janati El Idrissi, Y., Le Corff, S., and Moulines, E. (2024). Monte Carlo guided diffusion for Bayesian linear inverse problems. In *International Conference on Learning Representations*.

Villani, C. (2009). *Hypocoercivity*, volume 202 of *Memoirs of the American Mathematical Society*. American Mathematical Society.

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

Wu, L., Trippe, B., Naesseth, C., Blei, D., and Cunningham, J. P. (2023). Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pages 31372–31403.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., and Yang, M.-H. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39.

Supplementary Material for "Wasserstein Convergence of Critically Damped Langevin Diffusions"

Table of Contents

A Forward process of Critically-Damped dynamics

In this section, we establish several mathematical properties of the forward processes:

$$\mathrm{d}\overrightarrow{\mathbf{U}}_t = A\overrightarrow{\mathbf{U}}_t\mathrm{d}t + \Sigma_\varepsilon \mathrm{d}B_t\,, \qquad \overrightarrow{\mathbf{U}}_0 \sim \pi_{\mathrm{data}} \otimes \pi_v\,,$$

as defined in (4) with $\epsilon = 0$ or in (9) with $\epsilon \geq 0$. These results will be used throughout our subsequent analysis.

Lemma A.1. *Let A be the matrix defined in* (5)*, then*

$$A = \left(\begin{pmatrix} -a & -1 \\ 1 & 0 \end{pmatrix} \times \begin{pmatrix} -a & 1 \\ 0 & -a \end{pmatrix} \times \begin{pmatrix} 0 & 1 \\ -1 & -a \end{pmatrix}\right) \otimes \mathbf{I}_d$$

so that

$$\mathrm{e}^{tA} = \mathrm{e}^{-ta}\begin{pmatrix} 1+at & a^2t \\ -t & 1-at \end{pmatrix} \otimes \mathbf{I}_d\,, \tag{19}$$

and

$$\begin{aligned}\|\mathrm{e}^{tA}\| \leq \|\mathrm{e}^{tA}\|_1^{1/2}\|\mathrm{e}^{tA}\|_\infty^{1/2} &\leq (1+\max\{a+1; a(a+1)\}t)\,\mathrm{e}^{-ta} \\ &\leq \left(1+(a+1)^2t\right)\mathrm{e}^{-ta}\,.\end{aligned}$$

Proof. The Jordan matrix decomposition of A when $d = 1$ is given by,

$$A_1 = \begin{pmatrix} 0 & a^2 \\ -1 & -2a \end{pmatrix} = \begin{pmatrix} -a & -1 \\ 1 & 0 \end{pmatrix} \times \begin{pmatrix} -a & 1 \\ 0 & -a \end{pmatrix} \times \begin{pmatrix} 0 & 1 \\ -1 & -a \end{pmatrix}\,.$$

We can use this decomposition to obtain a matrix factorization in any dimension. As for all $k \in \mathbb{N}$, $A^k = (A_1^k \otimes \mathbf{I}_d)$,

$$\mathrm{e}^{tA} = \sum_{k=0}^{\infty} \frac{t^k}{k!}\left(A_1^k \otimes \mathbf{I}_d\right) = \left(\sum_{k=0}^{\infty}\frac{t^k}{k!}A_1^k\right) \otimes \mathbf{I}_d = \mathrm{e}^{tA_1} \otimes \mathbf{I}_d\,.$$

Finally, we deduce an upper bound to the spectral norm of e^{tA}, as

$$\|\mathrm{e}^{tA}\|_1 \leq \mathrm{e}^{-ta}\max\left\{(1+(a+1)t; 1+a(a+1)t\right\}\,,$$

and

$$\|\mathrm{e}^{tA}\|_\infty \leq \mathrm{e}^{-ta}\max\left\{(1+a(a+1)t; 1+(a+1)t\right\}\,.$$

Then,

$$\|\mathrm{e}^{tA}\| \leq \|\mathrm{e}^{tA}\|_1^{1/2}\|\mathrm{e}^{tA}\|_\infty^{1/2} \leq \mathrm{e}^{-ta}\left(1+\max\{a+1; a(a+1)\}t\right)\,,$$

which concludes the proof.

□

Lemma A.2. *Let $(\overrightarrow{\mathbf{U}}_t)_{t\in[0,T]}$ be a solution to the forward process* (9) *with initial condition*

$$\overrightarrow{\mathbf{U}}_0 \sim \pi_{\mathrm{data}} \otimes \pi_v\,,$$

where π_v is a probability distribution on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Then, the conditional law of $\overrightarrow{\mathbf{U}}_t$ given $\overrightarrow{\mathbf{U}}_0$, is Gaussian with mean $\mu_{t|0}$ and covariance $\Sigma_{0,t}$ defined by

$$\mu_{t|0} := \mathrm{e}^{tA}\overrightarrow{\mathbf{U}}_0\,, \qquad \Sigma_{0,t} := \Sigma_\infty - \mathrm{e}^{tA}\Sigma_\infty(\mathrm{e}^{tA})^\top, \tag{20}$$

with

$$\Sigma_\infty := \frac{1}{4}\begin{pmatrix} 5\varepsilon^2a^{-1} + a\sigma^2 & -2\varepsilon^2a^{-2} \\ -2\varepsilon^2a^{-2} & (\varepsilon^2 + a^2\sigma^2)a^{-3} \end{pmatrix} \otimes \mathbf{I}_d\,. \tag{21}$$

The result still holds when the forward process is defined as in (4) *by setting $\varepsilon = 0$.*

Proof. Recall that the forward process $(\overrightarrow{\mathbf{U}}_t)_{t\in[0,T]}$ is solution to,

$$\mathrm{d}\overrightarrow{\mathbf{U}}_t = A\overrightarrow{\mathbf{U}}_t\mathrm{d}t + \Sigma_\varepsilon \mathrm{d}B_t\,. \tag{22}$$

With initial condition $\overrightarrow{\mathbf{U}}_0 \sim \pi_{\mathrm{data}} \otimes \pi_v$, we have

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

This means that the law of $\overrightarrow{\mathbf{U}}_t$, conditional to the initial condition $\overrightarrow{\mathbf{U}}_0$ is Gaussian with mean

$$\mu_{t|0} := \mathbb{E}\left[\overrightarrow{\mathbf{U}}_t\right] = \mathrm{e}^{tA}\overrightarrow{\mathbf{U}}_0\,,$$

and covariance

$$\begin{aligned}\Sigma_{0,t} := \mathrm{Cov}\left(\overrightarrow{\mathbf{U}}_t\right) &= \int_0^t \mathrm{e}^{(t-s)A}\Sigma_\varepsilon^2(\mathrm{e}^{(t-s)A})^\top \mathrm{d}s \\ &= \int_0^t \mathrm{e}^{(t-s)A}\Sigma_\varepsilon^2(\mathrm{e}^{(t-s)A})^\top \mathrm{d}s \\ &= \left(\int_0^t \mathrm{e}^{(t-s)A_1}\Sigma_\varepsilon^2(\mathrm{e}^{(t-s)A_1})^\top \mathrm{d}s\right)\otimes \mathbf{I}_d\,.\end{aligned}$$

Using Lemma A.1, for $\delta \geq 0$,

$$\mathrm{e}^{\delta A_1}\Sigma_\varepsilon^2\mathrm{e}^{\delta A_1^\top} = \mathrm{e}^{-2a\delta}\begin{pmatrix} a^4\sigma^2\delta^2 + \varepsilon^2(1+a\delta)^2 & \delta\left(a^2\sigma^2\left(1-a\delta\right) - \varepsilon^2(1+a\delta)\right) \\ \delta\left(a^2\sigma^2\left(1-a\delta\right) - \varepsilon^2(1+a\delta)\right) & \sigma^2(1-a\delta)^2 + \delta^2\varepsilon^2\end{pmatrix}.$$

Hence, a straightforward computation provides with $\alpha_t = \left(-\left(5+2at(3+at)\right)\varepsilon^2 - a^2\left(1+2at(1+at)\right)\sigma^2\right)a^{-1}$ and $\gamma_t = 2\left((\varepsilon+at\varepsilon)^2 + a^4t^2\sigma^2\right)a^{-2}$,

$$\begin{aligned}\Sigma_{0,t} =& \frac{1}{4}\begin{pmatrix}5\varepsilon^2a^{-1} + a\sigma^2 & -2\varepsilon^2a^{-2} \\ -2\varepsilon^2a^{-2} & (\varepsilon^2+a^2\sigma^2)a^{-3}\end{pmatrix} \\ &+ \frac{\mathrm{e}^{-2at}}{4}\begin{pmatrix}\alpha_t & \gamma_t \\ \gamma_t & \left(-\left(1+2at(1+at)\right)\varepsilon^2 - a^2\left(1+2at(-1+at)\right)\sigma^2\right)a^{-3}\end{pmatrix} \\ =& \Sigma_\infty - \mathrm{e}^{tA}\Sigma_\infty(\mathrm{e}^{tA})^\top\,,\end{aligned} \tag{23}$$

where we used that,

$$\mathrm{e}^{tA}\Sigma_\infty(\mathrm{e}^{tA})^\top = \int_0^\infty \mathrm{e}^{(t+s)A}\Sigma^2\left(\mathrm{e}^{(t+s)A}\right)^\top \mathrm{d}s = \int_t^\infty \mathrm{e}^{\delta A}\Sigma^2\left(\mathrm{e}^{\delta A}\right)^\top \mathrm{d}\delta = \Sigma_\infty - \Sigma_{0,t}\,.$$

□

Lemma A.3. *The covariance matrix $\Sigma_{0,t}$ defined in* (20) *satisfies, for all $\varepsilon > 0$,*

$$\begin{aligned}\lambda_{\min}(\Sigma_{0,t}) \geq \max\Bigg\{&\frac{\sigma^2}{4}\min\{a,1/a\} - \left(\frac{\sigma^2}{4}\max\{a,1/a\} + \frac{5\varepsilon^2}{4a}\right)\mathrm{e}^{-2at}, \\ &\min\{\varepsilon^2,\sigma^2\}\frac{1-\mathrm{e}^{-2at}}{2a\big(1+(a+1)^2t\big)^2}\Bigg\}, \\ \lambda_{\max}(\Sigma_{0,t}) \leq& \frac{\sigma^2}{4}\max\{a,1/a\} + \frac{5\varepsilon^2}{4a}.\end{aligned}$$

Proof. First, consider the following decomposition of Σ_∞ defined in (21)

$$\Sigma_\infty = \frac{1}{4}\begin{pmatrix}a\sigma^2 & 0 \\ 0 & \sigma^2a^{-1}\end{pmatrix} + \frac{\varepsilon^2}{4a^3}\begin{pmatrix}5a^2 & -2a \\ -2a & 1\end{pmatrix} =: \frac{1}{4}\begin{pmatrix}a\sigma^2 & 0 \\ 0 & \sigma^2a^{-1}\end{pmatrix} + E_\varepsilon\,.$$

Since E_ε is positive definite, its trace and determinant are positive, then

$$\lambda_{\min}(\Sigma_\infty) \geq \frac{1}{4}\lambda_{\min}\left(\begin{pmatrix} a\sigma^2 & 0 \\ 0 & \sigma^2 a^{-1}\end{pmatrix}\right) = \frac{\sigma^2}{4}\min\{a, 1/a\}\,,$$
$$\lambda_{\max}(\Sigma_\infty) \leq \frac{1}{4}\lambda_{\max}\left(\begin{pmatrix} a\sigma^2 & 0 \\ 0 & \sigma^2 a^{-1}\end{pmatrix}\right) + \lambda_{\max}(E_\varepsilon) \leq \frac{\sigma^2}{4}\max\{a, 1/a\} + \frac{5\varepsilon^2}{4a}\,. \tag{24}$$

Using that $\Sigma_{0,t} = \Sigma_\infty - \mathrm{e}^{tA}\Sigma_\infty \mathrm{e}^{tA^\top}$ together with Weyl's inequality we have that

$$\lambda_{\min}(\Sigma_{0,t}) \geq \lambda_{\min}(\Sigma_\infty) - \lambda_{\max}\big(\mathrm{e}^{tA}\Sigma_\infty \mathrm{e}^{tA^\top}\big)\,.$$

Note that, as Σ_∞ is positive semidefinite,

$$\lambda_{\max}\big(\mathrm{e}^{tA}\Sigma_\infty \mathrm{e}^{tA^\top}\big) = \lambda_{\max}\big(\mathrm{e}^{tA}\Sigma_\infty^{1/2}\big)^2 \leq \lambda_{\max}\big(\mathrm{e}^{tA}\big)^2\lambda_{\max}\big(\Sigma_\infty\big) \leq \mathrm{e}^{-2at}\lambda_{\max}\big(\Sigma_\infty\big)\,.$$

On the other hand, using that $\Sigma_{0,t} = \int_0^t \mathrm{e}^{sA}\Sigma_\varepsilon^2 \mathrm{e}^{sA^\top}\mathrm{d}s$, yields

$$\Sigma_{0,t} \succcurlyeq \min\left\{\varepsilon^2, \sigma^2\right\}\int_0^t \mathrm{e}^{sA}\mathrm{e}^{sA^\top}\mathrm{d}s\,,$$

therefore,

$$\begin{aligned}
\lambda_{\min}(\Sigma_{0,t}) &\geq \min\left\{\varepsilon^2, \sigma^2\right\}\int_0^t \lambda_{\min}\left(\mathrm{e}^{sA}\mathrm{e}^{sA^\top}\right)\mathrm{d}s \\
&\geq \min\left\{\varepsilon^2, \sigma^2\right\}\int_0^t \frac{\mathrm{e}^{-2as}}{(1+(a+1)^2 s)^2}\mathrm{d}s, \\
&\geq \min\left\{\varepsilon^2, \sigma^2\right\}\frac{1-\mathrm{e}^{-2at}}{2a(1+(a+1)^2 t)^2}\,,
\end{aligned}$$

which gives the other lower bound of $\lambda_{\min}.(\Sigma_{0,t})$

To obtain the bound on $\lambda_{\max}(\Sigma_{0,t})$, it is enough to note that $\Sigma_{0,t} \preccurlyeq \Sigma_\infty$. □

Lemma A.4 (Forward process $\mathcal{W}_2$-contraction)**.** *The forward process, defined as in* (9)*, is contractive for the $\mathcal{W}_2$ distance. In particular, it holds that*

$$\mathcal{W}_2\left(\mathcal{L}(\overrightarrow{\mathbf{U}}_T), \pi_\infty\right) \leq K_T e^{-aT}\mathcal{W}_2\left(\pi_{\mathrm{data}}\otimes\pi_v, \pi_\infty\right)\,,$$

where π_∞ is the stationary distribution of (9) *as defined in Lemma A.2 and*

$$K_T := (1 + \max\{a+1; a(a+1)\}T)\,.$$

Proof. Let $u = (x, v) \in \mathbb{R}^{2d}$ (resp. $\bar{u} = (\bar{x}, \bar{v}) \in \mathbb{R}^{2d}$) and denote by $(\overrightarrow{\mathbf{U}}_t^u)_{t\in[0,T]}$ (resp. $(\overrightarrow{\mathbf{U}}_t^{\bar{u}})_{t\in[0,T]}$) the solution of (9), with initial condition $\overrightarrow{\mathbf{U}}_0^u = u$ (resp. $\overrightarrow{\mathbf{U}}_0^{\bar{u}} = \bar{u}$). By Itô's lemma,

$$\mathrm{d}\left(e^{-tA}\overrightarrow{\mathbf{U}}^{x,v}\right) = e^{-tA}\Sigma_\varepsilon \mathrm{d}B_t\,.$$

Using a synchronous coupling for $(\overrightarrow{\mathbf{U}}_t^u)_{t\in[0,T]}$ and $(\overrightarrow{\mathbf{U}}_t^{\bar{u}})_{t\in[0,T]}$, we have that

$$\overrightarrow{\mathbf{U}}_t^u - \overrightarrow{\mathbf{U}}_t^{\bar{u}} = \mathrm{e}^{tA}\left(u - \bar{u}\right)\,.$$

By definition of the Wasserstein-2 distance $\mathcal{W}_2(\mathcal{L}(\overrightarrow{\mathbf{U}}_t^u), \mathcal{L}(\overrightarrow{\mathbf{U}}_t^{\bar{u}})) \leq \|\overrightarrow{\mathbf{U}}_t^u - \overrightarrow{\mathbf{U}}_t^{\bar{u}}\|_{L_2}$. Then, by Lemma A.1,

$$\left\|\overrightarrow{\mathbf{U}}_t^u - \overrightarrow{\mathbf{U}}_t^{\bar{u}}\right\|_{L_2} \leq \left\|\mathrm{e}^{tA}\right\|\left\|u - \bar{u}\right\|_{L_2} \leq K_t \mathrm{e}^{-ta}\left\|u - \bar{u}\right\|_{L_2}\,, \tag{25}$$

with

$$K_t := (1 + \max\{a+1; a(a+1)\}t)\,.$$

Finally, assume that $\bar{u} \sim \pi_\infty$, $u \sim \pi_{\mathrm{data}}\otimes\pi_v$ and fix any coupling $\gamma \in \Pi(\pi_{\mathrm{data}}\otimes\pi_v, \pi_\infty)$. Using that π_∞ is stationary distribution of $\overrightarrow{\mathbf{U}}_t$ and taking the infimum over $\gamma \in \Pi(\pi_{\mathrm{data}}\otimes\pi_v, \pi_\infty)$ yields,

$$\mathcal{W}_2\left(\mathcal{L}\left(\overrightarrow{\mathbf{U}}_T\right), \pi_\infty\right) \leq K_T e^{-aT}\mathcal{W}_2\left(\pi_{\mathrm{data}}\otimes\pi_v, \pi_\infty\right)\,,$$

which finishes the proof. □

B Proof of Theorem 3.1

In this section we prove Theorem 3.1. We use notations from (12) (resp. (13)) for the continuous time interpolation of the discretized backward with modified score function $\bar{\mathbf{U}}_t$ (resp. for the continuous time interpolation of the discretized backward with approximated modified score function $\bar{\mathbf{U}}_t^\theta$). We first establish the propagation of Lipschitz regularity, followed by the proof of Theorem 3.1. To do so, we decompose the generation error as the sum of the discretization error (Lemma B.2), the approximation error (Lemma B.3), and the mixing time error (Lemma B.4).

B.1 Propagation of the regularity assumptions

Proposition B.1. *Assume that Assumption H2 holds. Then, for all $t > 0$, $\Sigma_\varepsilon^2 \nabla \log p_t$ (resp. $\Sigma_\varepsilon^2 \nabla \log \tilde{p}_t$) is L_t-Lipschitz (resp. $\tilde{L}_t$-Lipschitz): for all $u \in \mathbb{R}^{2d}$,*

$$\left\|\Sigma_\varepsilon^2 \nabla^2 \log p_t(u)\right\| \leq L_t \,.$$

Moreover, there exists a constant $C > 0$ such that for all $u \in \mathbb{R}^{2d}$,

$$\left\|\Sigma_\varepsilon^2 \nabla^2 \log \tilde{p}_t(u)\right\| \leq \tilde{L}_t \leq C\left(1 + \frac{1}{\sqrt{t}}\right) \mathrm{e}^{-2at} \,. \tag{26}$$

Proof. Step 1: Lower bound on $\nabla^2 \log p_t$. Recall the following equality in law given by the modified kinetic OU process (9)

$$\overrightarrow{\mathbf{U}}_t \overset{\mathcal{L}}{=} \mathrm{e}^{tA}\overrightarrow{\mathbf{U}}_0 + \sqrt{\Sigma_{0,t}}G \,,$$

with $\overrightarrow{\mathbf{U}}_0 \sim \pi_{\mathrm{data}} \otimes \pi_v$, $G \sim \mathcal{N}(0, \mathbf{I}_{2d})$, where G and $\overrightarrow{\mathbf{U}}_0$ are independent, and $\Sigma_{0,t}$ is defined in (20). Writing $q_{t|0}$ the conditional density of $\overrightarrow{\mathbf{U}}_t$ given $\overrightarrow{\mathbf{U}}_0$, we have

$$\begin{aligned}
p_t(u_t) &= \int_{\mathbb{R}^{2d}} p_0(u_0) q_{t|0}(u_t|u_0) \mathrm{d}u_0 \\
&= \int_{\mathbb{R}^{2d}} p_0(u_0)\, \det\left(2\pi\Sigma_{0,t}\right)^{-1/2} \exp\left(-\frac{1}{2}\left(u_t - \mathrm{e}^{tA}u_0\right)^\top \Sigma_{0,t}^{-1}\left(u_t - \mathrm{e}^{tA}u_0\right)\right) \mathrm{d}u_0 \\
&= \det\left(\mathrm{e}^{-tA}\right) \int_{\mathbb{R}^{2d}} p_0\left(\mathrm{e}^{-tA}z\right) \det\left(2\pi\Sigma_{0,t}\right)^{-1/2} \exp\left(-\frac{1}{2}\Big(u_t - z\Big)^\top \Sigma_{0,t}^{-1}\Big(u_t - z\Big)\right) \mathrm{d}z.
\end{aligned}$$

As also observed in Saumard and Wellner (Proposition 7.1, 2014), we get

$$\begin{aligned}
\nabla^2 \log p_t(u) &= \mathrm{Var}(\nabla\phi_{0,t}(Y_0)|Y_0 + Y_1 = u) - \mathbb{E}[\nabla^2\phi_{0,t}(Y_0)|Y_0 + Y_1 = u] \\
&= \mathrm{Var}(\nabla\phi_{1,t}(Y_1)|Y_0 + Y_1 = u) - \mathbb{E}[\nabla^2\phi_{1,t}(Y_1)|Y_0 + Y_1 = u] \,,
\end{aligned} \tag{27}$$

for $Y_0 = \mathrm{e}^{tA}\overrightarrow{\mathbf{U}}_0$ and $Y_1 = \sqrt{\Sigma_{0,t}}G$ and for $\phi_{0,t}$ and $\phi_{1,t}$ such that for all $u \in \mathbb{R}^{2d}$,

$$\begin{aligned}
\mathrm{e}^{-\phi_{0,t}(u)} &:= \det\left(\mathrm{e}^{-tA}\right) p_0\left(\mathrm{e}^{-tA}u\right) \,, \\
\mathrm{e}^{-\phi_{1,t}(u)} &:= \det\left(2\pi\Sigma_{0,t}\right)^{-1/2} \exp\left(-\frac{1}{2}u^\top\Sigma_{0,t}^{-1}u\right) \,.
\end{aligned}$$

This implies that

$$\begin{aligned}
\nabla^2 \log p_t(u) &\succcurlyeq -\mathbb{E}[\nabla^2\phi_{0,t}(Y_0)|Y_0 + Y_1 = u] \,, \\
\nabla^2 \log p_t(u) &\succcurlyeq -\mathbb{E}[\nabla^2\phi_{1,t}(Y_1)|Y_0 + Y_1 = u] \,.
\end{aligned} \tag{28}$$

Note that for all $u \in \mathbb{R}^{2d}$,

$$\begin{aligned}
\nabla^2\phi_{0,t}(u) &= -\mathrm{e}^{-tA^\top} \nabla^2 \log p_0\left(\mathrm{e}^{-tA}u\right) \mathrm{e}^{-tA} \,, \\
\nabla^2\phi_{1,t}(u) &= \Sigma_{0,t}^{-1} \,.
\end{aligned}$$

From (Bouchut et al., 2005, Lemma 2.2) together with (14), we get that the one-sided Lipschitz assumption entails the following inequality over the Hessian of the log-density, since $\log p_0(u) = \log \pi_{\mathrm{data}}(x) + \log p_v(v)$,

$$\begin{aligned}\nabla^2 \left(-\log p_0\right)(u) &= \begin{pmatrix} -\nabla^2 \log p_{\mathrm{data}}(x) & 0 \\ 0 & -\nabla^2 \log p_v(v) \end{pmatrix} \\ &= \begin{pmatrix} -\nabla^2 \log p_{\mathrm{data}}(x) & 0 \\ 0 & v^{-2}\mathbf{I}_d \end{pmatrix} \preccurlyeq \max\left\{L_0, \frac{1}{v^2}\right\} \mathbf{I}_{2d}.\end{aligned}$$

Therefore, for $t > 0$, from (28), we get

$$\nabla^2 \log p_t(u) \succcurlyeq -\mathfrak{h}_t \mathbf{I}_{2d}\,,$$

where $\mathfrak{h}_t = \min\left\{ \left\|\mathrm{e}^{-tA}\right\|^2 \max\left\{L_0, v^{-2}\right\}; \left\|\Sigma_{0,t}^{-1}\right\| \right\}$.

Bound on $\mathfrak{h}_t$. By Lemma A.1, we have that $\left\|\mathrm{e}^{-tA}\right\|^2 \leq \left(1 + (a+1)^2 t\right)^2 \mathrm{e}^{2ta}$. Moreover, from Lemma A.3, it follows that

$$\left\|\Sigma_{0,t}^{-1}\right\| = \frac{1}{\lambda_{\min}(\Sigma_{0,t})} \leq \frac{1}{\lfloor \lambda_{\min}(\Sigma_\infty) - \lambda_{\max}(\mathrm{e}^{tA}\Sigma_\infty \mathrm{e}^{tA^\top}) \rfloor_+}\,,$$

with $\lfloor \cdot \rfloor_+$ denoting the positive part of a real number. Therefore,

$$\left\|\Sigma_{0,t}^{-1}\right\| \leq \frac{4}{\lfloor \sigma^2 \min\{a, 1/a\} - (\sigma^2 \max\{a, 1/a\} + 5\varepsilon^2 a^{-1}) \mathrm{e}^{-2at} \rfloor_+} =: \mathfrak{h}_{2,t}\,.$$

Combining the two previous bounds, we obtain

$$\mathfrak{h}_t \leq \min\left\{\mathfrak{h}_{1,t}; \mathfrak{h}_{2,t}\right\}\,, \tag{29}$$

where $\mathfrak{h}_{1,t} := \left(1 + (a+1)^2 t\right)^2 \mathrm{e}^{2ta} \max\left\{L_0, v^{-2}\right\}$.

Step 2: Upper bound on $\nabla^2 \log p_t$. We first express the conditional density of $\overrightarrow{\mathbf{U}}_0$ given $\overrightarrow{\mathbf{U}}_t$ as follows

$$q_{t|0}((x_0, v_0)^\top | u_t) \propto \left(\mathrm{e}^{-V(x_0) - H(x_0)} \otimes \mathcal{N}(v_0; 0_d, v^2 \mathbf{I}_d)\right) \mathcal{N}(u_t; e^{tA}(x_0, v_0)^\top, \Sigma_{0,t})\,. \tag{30}$$

First, we consider the log-concave part of the above distribution,

$$\nu_t \propto \left(\mathrm{e}^{-V(x_0)} \otimes \mathcal{N}(v_0; 0_d, v^2 \mathbf{I}_d)\right) \mathcal{N}(u_t; e^{tA}(x_0, v_0)^\top, \Sigma_{0,t})\,. \tag{31}$$

Since $\nabla^2 V(x) \succcurlyeq \alpha \mathbf{I}_d$ for all $x \in \mathbb{R}^d$, we obtain

$$\nabla^2 \left(-\log \nu_t\right) \succcurlyeq \mathrm{e}^{-tA} \begin{pmatrix} \alpha \mathbf{I}_d & 0 \\ 0 & \frac{1}{v^2}\mathbf{I}_d \end{pmatrix} \mathrm{e}^{-tA^\top} + \Sigma_{0,t}^{-1}\,.$$

Therefore, by Brascamp–Lieb inequality (Brascamp and Lieb, 1976),

$$\mathrm{Cov}(\nu_t) \preccurlyeq \left(\mathrm{e}^{-tA} \begin{pmatrix} \alpha \mathbf{I}_d & 0 \\ 0 & \frac{1}{v^2}\mathbf{I}_d \end{pmatrix} \mathrm{e}^{-tA^\top} + \Sigma_{0,t}^{-1}\right)^{-1}\,.$$

Using the identity $\Sigma_{0,t} = \Sigma_\infty - e^{tA}\Sigma_\infty e^{tA^T}$ given in Lemma A.2, we now expand $\Sigma_{0,t}$ at zero as

$$\Sigma_{0,t} = t \begin{pmatrix} \varepsilon^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} + \mathcal{O}(t^2)\,,$$

which implies that

$$\Sigma_{0,t}^{-1} = \frac{1}{t} \underbrace{\begin{pmatrix} 1/\varepsilon^2 & 0 \\ 0 & 1/\sigma^2 \end{pmatrix}}_{=\Sigma_\varepsilon^{-1}} + o\left(\tfrac{1}{t}\right)\,.$$

Therefore, the covariance matrix near zero satisfies

$$\mathrm{Cov}(\nu_t) \preccurlyeq \begin{pmatrix} \left(\alpha + \frac{1}{\varepsilon^2 t}\right)^{-1} & 0 \\ 0 & \left(\frac{1}{v^2} + \frac{1}{\sigma^2 t}\right)^{-1} \end{pmatrix} + o(t)\,.$$

Next, the Lipschitz perturbation term, following Brigati and Pedrotti (2025), can be bounded as

$$\mathrm{Cov}(q_t(.|u_t)) \preccurlyeq \underbrace{\begin{pmatrix} \left(\frac{L}{\alpha+(\epsilon^2 t)^{-1}} + \sqrt{\frac{1}{\alpha+(\epsilon^2 t)^{-1}}}\right)^2 & 0 \\ 0 & \left(\frac{1}{v^2}+\frac{1}{\sigma^2 t}\right)^{-1} \end{pmatrix}}_{:=M_{\varepsilon,t}} + o(t)\,.$$

Using (27), we have

$$\nabla^2 \log p_t(u) = \Sigma_{0,t}^{-1}\mathrm{Cov}(q_t(.|u))\Sigma_{0,t}^{-1} - \Sigma_{0,t}^{-1}\,, \tag{32}$$

so that

$$\begin{aligned}\nabla^2 \log p_t(u) &= \left(\frac{1}{t}\Sigma_\varepsilon^{-1} + o\left(\frac{1}{t}\right)\right)(M_t + o(t))\left(\frac{1}{t}\Sigma_\varepsilon^{-1} + o\left(\frac{1}{t}\right)\right) - \left(\frac{1}{t}\Sigma_\varepsilon^{-1} + o\left(\frac{1}{t}\right)\right) \\ &= \begin{pmatrix}\alpha_t & 0\\ 0 & \beta_t\end{pmatrix} + o\left(\frac{1}{t}\right)\,,\end{aligned}$$

with

$$|\alpha_t| \le \frac{L^2}{(\alpha\epsilon^2 t+1)^2} + \frac{2L}{(\epsilon^2 t)^{1/2}(\alpha\epsilon^2 t+1)^{3/2}} - \frac{\alpha}{\alpha\epsilon^2 t+1}, \qquad \beta_t := -\frac{1}{\sigma^2 t + v^2}\,.$$

Consequently, for all $\epsilon > 0$, as $t \to 0^+$,

$$\begin{pmatrix}\frac{2L}{\epsilon\sqrt{t}} & 0\\ 0 & \frac{1}{v^2}\end{pmatrix} + o\left(\frac{1}{\sqrt{t}}\right) \le \nabla^2 \log p_t(u) \le \begin{pmatrix}\frac{2L}{\epsilon\sqrt{t}} & 0\\ 0 & -\frac{1}{v^2}\end{pmatrix} + o\left(\frac{1}{\sqrt{t}}\right)\,. \tag{33}$$

Step 3: Uniform bound on $\nabla^2 \log p_t$. We now analyze the structure of the minimum in the upper bound of $\mathfrak{h}_t$ in (29). We observe that the first term is increasing, equals $\max\left\{L_0, v^{-2}\right\}$ for $t \to 0$, and diverges as $t \to +\infty$. In contrast, the second term is decreasing: it diverges as $t \to 0$ and converges to $4/(\sigma^2 \min\{a, 1/a\})$ as $t \to +\infty$. Therefore, the minimum coincides with the first term, for $t \le T_{\mathrm{change}}$, and with the second term, for $t > T_{\mathrm{change}}$. Using (33), we obtain the following uniform bound, for all $\epsilon > 0$,

$$\left\|\nabla^2 \log p_t(u)\right\| \le \max\left\{\mathfrak{h}_{1,T_{\mathrm{change}}}; Ct^{-1/2}\right\}\,, \quad \text{for } t > 0\,. \tag{34}$$

This bound is uniform in $\varepsilon > 0$, therefore, for $\varepsilon \to 0$, we have

$$\left\|\Sigma_\epsilon^2 \nabla^2 \log p_t(u)\right\| \le \max\left\{\mathfrak{h}_{1,T_{\mathrm{change}}}; Ct^{-1/2}\right\}\,, \quad \text{for } t > 0\,. \tag{35}$$

Since $\tilde{p}_t = p_t/p_\infty$, and p_∞ is the density of a centered Gaussian vector of variance Σ_∞, we have

$$\nabla^2 \log \tilde{p}_t(u) = \nabla^2 \log p_t(u) + \Sigma_\infty^{-1}\,. \tag{36}$$

Therefore, the same bound as in (35) holds for the modified score.

Step 4: Exponential decay of the modified score. From (32), we have the following equality

$$\nabla^2 \log \tilde{p}_t(u) = \nabla^2 \log p_t(u) + \Sigma_\infty^{-1} = \Sigma_{0,t}^{-1}\mathrm{Cov}(q_t(.|u))\Sigma_{0,t}^{-1} - \Sigma_{0,t}^{-1} + \Sigma_\infty^{-1}\,,$$

where q_t is defined in (30). By applying Lemma D.9, together with the decomposition (20) and the positivity of the covariance, we obtain

$$\nabla^2 \log \tilde{p}_t(u) \succcurlyeq -\Sigma_{0,t}^{-1}\left(\mathrm{e}^{tA}\Sigma_\infty(\mathrm{e}^{tA})^\top\right)\Sigma_\infty^{-1}\,.$$

Since $\Sigma_{0,t} = \Sigma_\infty + \mathcal{O}(e^{-2at})$ as $t \to \infty$, there exists a constant $C > 0$ such that

$$\left\|\Sigma_{0,t}^{-1}\left(\mathrm{e}^{tA}\Sigma_\infty(\mathrm{e}^{tA})^\top\right)\Sigma_\infty^{-1}\right\| \le Ce^{-2at}$$

On the other hand, using the fact that $\Sigma_{0,t}^{-1} \succcurlyeq \Sigma_\infty^{-1}$ (see (20)), we get

$$\nabla^2 \log \tilde{p}_t(u) \preccurlyeq \Sigma_{0,t}^{-1}\mathrm{Cov}(q_t(.|u))\Sigma_{0,t}^{-1}\,.$$

Following the same steps as in the derivation of the upper bound on $\nabla^2 \log p_t$ (step 2), we obtain

$$\nabla^2 (-\log \nu_t) \succcurlyeq \mathrm{e}^{-tA} \begin{pmatrix} \alpha \mathbf{I}_d & 0 \\ 0 & \frac{1}{v^2}\mathbf{I}_d \end{pmatrix} \mathrm{e}^{-tA^\top} + \Sigma_{0,t}^{-1} \succcurlyeq \mathrm{e}^{-tA} \begin{pmatrix} \alpha \mathbf{I}_d & 0 \\ 0 & \frac{1}{v^2}\mathbf{I}_d \end{pmatrix} \mathrm{e}^{-tA^\top} ,$$

where ν_t is defined in (30). By the Brascamp–Lieb inequality, this implies that $\mathrm{Cov}(\nu_t) = \mathcal{O}(e^{-2at})$. Next, similarly to "step 2", for the term involving the Lipschitz perturbation, and following Brigati and Pedrotti (2025, Theorem 1.3), we have

$$\mathrm{Cov}(q_t(.|u)) \preccurlyeq \left(LCe^{-2at} + \sqrt{Ce^{-2at}} \right)^2 \mathbf{I}_d \, .$$

Therefore, there exist a universal constant $C > 0$ and a finite time $T_{\mathrm{change}} > 0$ such that, for all $t \geq T_{\mathrm{change}}$,

$$\left\| \nabla^2 \log \tilde{p}_t(u) \right\| \leq C\mathrm{e}^{-2at} \, .$$

This implies that the modified score function is $\tilde{L}_t$-Lipschitz, with $\tilde{L}_t$ defined as

$$\tilde{L}_t := \begin{cases} \max\left\{ \mathfrak{h}_{1,T_{\mathrm{change}}}; Ct^{-1/2} \right\} + \max\{a, 1/a\} \, , & \text{for } t \in (0, T_{\mathrm{change}}] \, , \\ C\mathrm{e}^{-2at} \, , & \text{for } t \in (T_{\mathrm{change}}, +\infty) \, , \end{cases}$$

which concludes the proof.

□

B.2 Proofs of the main results

Lemma B.2 (Discretization error)**.** *Assume that H1 and H2 hold. Then, for all $\eta > 0$ and all $h > 0$, there exists a constant $C > 0$ such that*

$$\mathcal{W}_2 \left(\mathcal{L}(\overleftarrow{\mathbf{U}}_T), \mathcal{L}(\bar{\mathbf{U}}_T) \right) \leq \sqrt{h} C \sqrt{ \left(h\|A\|^2 B^2 + \|\Sigma_\epsilon\|^2 \left(d + \mathcal{I}(\pi_{\mathrm{data}} \otimes \pi_v | \pi_\infty) \right) \right) \frac{\mathrm{e}^{Ca^{-1}}}{a - \eta} } \, , \quad (37)$$

with

$$B := \max_{t \in [0,T]} \left(1 + (a+1)^2 (T-t) \right)^2 \mathrm{e}^{-2a(T-t)} \left\| \overrightarrow{\mathbf{U}}_0 \right\|_{L_2}^2 + \frac{d}{2} \left(\sigma^2 \max\{a, 1/a\} + \frac{5\varepsilon^2}{a} \right) .$$

Proof. Consider a synchronous coupling for $(\overleftarrow{\mathbf{U}}_t)_{t \in [0,T]}$ and $(\bar{\mathbf{U}}_t)_{t \in [0,T]}$, *i.e.*, use the same Brownian motion to drive the two processes, with the same initial point, *i.e.*, $\overleftarrow{\mathbf{U}}_0 = \bar{\mathbf{U}}_0$. Then, it holds that

$$\mathcal{W}_2 \left(\mathcal{L}(\overleftarrow{\mathbf{U}}_T), \mathcal{L}(\bar{\mathbf{U}}_T) \right) \leq \left\| \overleftarrow{\mathbf{U}}_T - \bar{\mathbf{U}}_T \right\|_{L_2} .$$

Fix $0 < \Delta < h$ and let $t_N := T - \Delta$. Note that, for all $0 \leq k \leq N-1$, from (11) and (12),

$$\begin{aligned} &\overleftarrow{\mathbf{U}}_{t_{k+1}} - \bar{\mathbf{U}}_{t_{k+1}} \\ &\qquad = \overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k} + \int_{t_k}^{t_{k+1}} \left\{ \tilde{A}_\epsilon \left(\overleftarrow{\mathbf{U}}_t - \bar{\mathbf{U}}_{t_k} \right) + \Sigma_\epsilon^2 \left(\tilde{\mathsf{s}}_{T-t} \left(\overleftarrow{\mathbf{U}}_t \right) - \tilde{\mathsf{s}}_{T-t_k} \left(\bar{\mathbf{U}}_{t_k} \right) \right) \right\} \mathrm{d}t \, , \end{aligned}$$

where

$$\tilde{A}_\epsilon = -A - \Sigma_\epsilon^2 \Sigma_\infty^{-1} \, .$$

From Monmarché (2023, Lemma 5 and Proposition 4) and Achleitner et al. (2015, Lemma 2.6), there exists a symmetric positive definite matrix $\mathfrak{M} \in \mathbb{R}^{2d \times 2d}$ such that, for any fixed $\eta > 0$, we have

$$\mathfrak{M} \tilde{A}_\epsilon \preccurlyeq -(a - \eta) \mathfrak{M} \, . \quad (38)$$

We then prove contraction with respect to the norm associated with $\mathfrak{M}$ defined, for all $v \in \mathbb{R}^{2d}$, by $\|v\|_{\mathfrak{M}}^2 := v^\top \mathfrak{M} v$.

For $t \in [t_k, t_{k+1})$,

$$\mathrm{d} \left(\overleftarrow{\mathbf{U}}_t - \bar{\mathbf{U}}_t \right) = \tilde{A}_\epsilon \left(\overleftarrow{\mathbf{U}}_t - \bar{\mathbf{U}}_{t_k} \right) \mathrm{d}t + \Sigma_\epsilon^2 \left(\tilde{s}_{T-t} \left(\overleftarrow{\mathbf{U}}_t \right) - \tilde{s}_{T-t_k} \left(\bar{\mathbf{U}}_{t_k} \right) \right) \mathrm{d}t \, .$$

This means that we have

$$\mathrm{d}\left(\left(\overleftarrow{\mathbf{U}}_t-\bar{\mathbf{U}}_t\right)^\top\mathfrak{M}\left(\overleftarrow{\mathbf{U}}_t-\bar{\mathbf{U}}_t\right)\right)=2\left(\overleftarrow{\mathbf{U}}_t-\bar{\mathbf{U}}_t\right)^\top\mathfrak{M}\mathrm{d}\left(\overleftarrow{\mathbf{U}}_t-\bar{\mathbf{U}}_t\right).$$

It follows that,

$$\begin{aligned}\left\|\overleftarrow{\mathbf{U}}_{t_{k+1}}-\bar{\mathbf{U}}_{t_{k+1}}\right\|_{\mathfrak{M}}^2=&\left\|\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2+2\int_{t_k}^{t_{k+1}}\left(\overleftarrow{\mathbf{U}}_t-\bar{\mathbf{U}}_{t_k}\right)^\top\mathfrak{M}\tilde{A}_\epsilon\left(\overleftarrow{\mathbf{U}}_t-\bar{\mathbf{U}}_{t_k}\right)\mathrm{d}t\\&+2\int_{t_k}^{t_{k+1}}\left(\overleftarrow{\mathbf{U}}_t-\bar{\mathbf{U}}_{t_k}\right)^\top\mathfrak{M}\Sigma_\epsilon^2\left(\tilde{\mathsf{s}}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)-\tilde{\mathsf{s}}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\mathrm{d}t\\\leq&\left\|\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2+2\left(A_{1,k}+A_{2,k}+A_{3,k}+A_{4,k}+A_{5,k}+A_{6,k}\right),\end{aligned}$$

where

$$\begin{aligned}A_{1,k}:=&h\left(\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right)^\top\mathfrak{M}\tilde{A}_\epsilon\left(\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right)\\&+h\left(\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right)^\top\mathfrak{M}\Sigma_\epsilon^2\left(\tilde{\mathsf{s}}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)-\tilde{\mathsf{s}}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right),\\A_{2,k}:=&\int_{t_k}^{t_{k+1}}\left(\overleftarrow{\mathbf{U}}_t-\overleftarrow{\mathbf{U}}_{t_k}\right)^\top\mathfrak{M}\left\{\tilde{A}_\epsilon\left(\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right)+\Sigma_\epsilon^2\left(\tilde{\mathsf{s}}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)-\tilde{\mathsf{s}}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\right\}\mathrm{d}t,\\A_{3,k}:=&\int_{t_k}^{t_{k+1}}\left(\overleftarrow{\mathbf{U}}_t-\overleftarrow{\mathbf{U}}_{t_k}\right)^\top\tilde{A}_\epsilon^\top\mathfrak{M}\left(\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right)\mathrm{d}t,\\A_{4,k}:=&\left(\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right)^\top\mathfrak{M}\Sigma_\epsilon^2\int_{t_k}^{t_{k+1}}\left(\tilde{\mathsf{s}}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)-\tilde{\mathsf{s}}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)\right)\mathrm{d}t,\\A_{5,k}:=&\int_{t_k}^{t_{k+1}}\left(\overleftarrow{\mathbf{U}}_t-\overleftarrow{\mathbf{U}}_{t_k}\right)^\top\mathfrak{M}\tilde{A}_\epsilon\left(\overleftarrow{\mathbf{U}}_t-\overleftarrow{\mathbf{U}}_{t_k}\right)\mathrm{d}t,\\A_{6,k}:=&\int_{t_k}^{t_{k+1}}\left(\overleftarrow{\mathbf{U}}_t-\overleftarrow{\mathbf{U}}_{t_k}\right)^\top\mathfrak{M}\Sigma_\epsilon^2\left(\tilde{\mathsf{s}}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)-\tilde{\mathsf{s}}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)\right)\mathrm{d}t.\end{aligned}$$

Next, we bound each term of the above decomposition separately.

Bound of $\mathbb{E}[A_{1,k}]$. By Assumption H2, applying Proposition B.1, there exists a constant C (that depends on the eigenvalues of $\mathfrak{M}$ or constant terms and that may vary from line to line) such that

$$\left(\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right)^\top\mathfrak{M}\Sigma_\epsilon^2\left(\tilde{\mathsf{s}}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)-\tilde{\mathsf{s}}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\leq C\tilde{L}_{T-t_k}\left\|\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2,$$

and using (38),

$$\left(\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right)^\top\mathfrak{M}\tilde{A}_\epsilon\left(\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right)\leq-(a-\eta)\left\|\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2.$$

Combining this with (38) yields

$$\mathbb{E}\left[A_{1,k}\right]\leq h\left(C\tilde{L}_{T-t_k}-(a-\eta)\right)\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right].$$

Bound of $\mathbb{E}[A_{2,k}]$. Using the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}\left[A_{2,k}\right]\leq&\mathbb{E}\left[\left\|\int_{t_k}^{t_{k+1}}\left(\overleftarrow{\mathbf{U}}_t-\overleftarrow{\mathbf{U}}_{t_k}\right)\mathrm{d}t\right\|^2\right]^{1/2}\\&\times\mathbb{E}\left[\left\|\mathfrak{M}\left\{\tilde{A}_\epsilon\left(\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right)+\Sigma_\epsilon^2\left(\tilde{\mathsf{s}}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)-\tilde{\mathsf{s}}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\right\}\right\|^2\right]^{1/2}\\\leq&C\mathbb{E}\left[\left\|\int_{t_k}^{t_{k+1}}\left(\overleftarrow{\mathbf{U}}_t-\overleftarrow{\mathbf{U}}_{t_k}\right)\mathrm{d}t\right\|^2\right]^{1/2}\mathbb{E}\left[\left\|\sqrt{\mathfrak{M}}\left(\tilde{b}_{t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)-\tilde{b}_{t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\right\|^2\right]^{1/2},\end{aligned}$$

with $\tilde{b}_t$ the backward drift in (11) defined by $\tilde{b}_t : u \mapsto \tilde{A}_\epsilon u + \Sigma_\epsilon^2 \tilde{\mathfrak{s}}_{T-t}(u)$. On the one hand, the Cauchy-Schwarz inequality implies

$$\mathbb{E}\left[\left\|\int_{t_k}^{t_{k+1}} \left(\overleftarrow{\mathbf{U}}_t - \overleftarrow{\mathbf{U}}_{t_k}\right) \mathrm{d}t\right\|^2\right]^{1/2} \leq \sqrt{h}\left(\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_t - \overleftarrow{\mathbf{U}}_{t_k}\right\|^2\right] \mathrm{d}t\right)^{1/2} .$$

Using the time-reversal property, Lemma D.3, together with Cauchy-Schwarz inequality and Itô's isometry,

$$\begin{aligned}
\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_t - \overleftarrow{\mathbf{U}}_{t_k}\right\|^2\right] &\leq \mathbb{E}\left[\left\|\int_{T-t}^{T-t_k} A\overrightarrow{\mathbf{U}}_s \mathrm{d}s + \Sigma_\epsilon \mathrm{d}B_s\right\|^2\right] \\
&\leq C\left(h\|A\|^2 \int_{T-t}^{T-t_k} \mathbb{E}\left[\left\|\overrightarrow{\mathbf{U}}_s\right\|^2\right] \mathrm{d}s + hd\left\|\Sigma_\epsilon\right\|^2\right) \\
&\leq C\left(h^2\|A\|^2 B^2 + hd\left\|\Sigma_\epsilon\right\|^2\right) ,
\end{aligned} \tag{39}$$

where B is defined in (56). It follows that

$$\mathbb{E}\left[\left\|\int_{t_k}^{t_{k+1}} \left(\overleftarrow{\mathbf{U}}_t - \overleftarrow{\mathbf{U}}_{t_k}\right) \mathrm{d}t\right\|^2\right]^{1/2} \leq h\sqrt{h}C\left(h\|A\|^2 B^2 + \left\|\Sigma_\epsilon\right\|^2 d\right)^{1/2} .$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sqrt{\mathfrak{M}}\left(\tilde{b}_{t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right) - \tilde{b}_{t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\right\|^2\right]^{1/2} \\
&\qquad\qquad \leq C\left(\|\tilde{A}_\epsilon\| + \tilde{L}_{T-t_k}\right)\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right]^{1/2} .
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[A_{2,k}\right] &\leq Ch\sqrt{h}\left(\|\tilde{A}_\epsilon\| + \tilde{L}_{T-t_k}\right)\left(h\|A\|^2 B^2 + \left\|\Sigma_\epsilon\right\|^2 d\right)^{1/2} \mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right]^{1/2} \\
&= Ch\sqrt{h}\|\tilde{A}_\epsilon\|\left(h\|A\|^2 B^2 + \left\|\Sigma_\epsilon\right\|^2 d\right)^{1/2} \mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right]^{1/2} \\
&\qquad + Ch\sqrt{h}\tilde{L}_{T-t_k}\left(h\|A\|^2 B^2 + \left\|\Sigma_\epsilon\right\|^2 d\right)^{1/2} \mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right]^{1/2} .
\end{aligned}$$

Moreover, from Young's inequality, we get that, for all $a, b \geq 0$ and $\alpha > 0$,

$$ab \leq \frac{\alpha}{2}a^2 + \frac{1}{2\alpha}b^2 . \tag{40}$$

It follows that,

$$\begin{aligned}
\mathbb{E}\left[A_{2,k}\right] \leq \frac{a-\eta}{6} h\,\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right] + h^2 C\frac{\|\tilde{A}_\epsilon\|^2\left(h\|A\|^2 B^2 + \left\|\Sigma_\epsilon\right\|^2 d\right)}{a-\eta} \\
+ Ch\tilde{L}_{T-t_k}\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right] + Ch^2\tilde{L}_{T-t_k}\left(h\|A\|^2 B^2 + \left\|\Sigma_\epsilon\right\|^2 d\right) .
\end{aligned}$$

Bound of $\mathbb{E}[A_{3,k}]$. Using Cauchy-Schwarz inequality,

$$\mathbb{E}\left[A_{3,k}\right] \leq \mathbb{E}\left[\left\|\int_{t_k}^{t_{k+1}} \left(\overleftarrow{\mathbf{U}}_t - \overleftarrow{\mathbf{U}}_{t_k}\right) \mathrm{d}t\right\|^2\right]^{1/2} \mathbb{E}\left[\left\|\tilde{A}_\epsilon^\top \mathfrak{M}\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right)\right\|^2\right]^{1/2} .$$

On the one hand,

$$\mathbb{E}\left[\left\|\tilde{A}_\epsilon^\top \mathfrak{M}\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right)\right\|^2\right]^{1/2} \leq C\|\tilde{A}_\epsilon\|\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right\|^2\right]^{1/2} ,$$

and, on the other hand, using (39) yields,

$$\mathbb{E}\left[\left\|\int_{t_k}^{t_{k+1}}\left(\overleftarrow{\mathbf{U}}_t-\overleftarrow{\mathbf{U}}_{t_k}\right)\mathrm{d}t\right\|^2\right]^{1/2}\leq Ch\sqrt{h}\left(h\|A\|^2B^2+\|\Sigma_\epsilon\|^2 d\right)^{1/2}.$$

Combining both and using (40) yield,

$$\begin{aligned}\mathbb{E}\left[A_{3,k}\right]&\leq Ch\sqrt{h}\|\tilde{A}_\epsilon\|\sqrt{h\|A\|^2B^2+\|\Sigma_\epsilon\|^2 d}\times\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right]^{1/2}\\&\leq\frac{a-\eta}{6}h\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right]+h^2C\frac{\|\tilde{A}_\epsilon\|^2\left(h\|A\|^2B^2+\|\Sigma_\epsilon\|^2 d\right)}{a-\eta}.\end{aligned}$$

Bound of $\mathbb{E}[A_{4,k}]$. Using Cauchy-Schwarz inequality,

$$\mathbb{E}\left[A_{4,k}\right]\leq C\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right]^{1/2}\mathbb{E}\left[\left\|\int_{t_k}^{t_{k+1}}\Sigma_\epsilon^2\left(\tilde{\mathfrak{s}}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)-\tilde{\mathfrak{s}}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)\right)\mathrm{d}t\right\|^2\right]^{1/2}.$$

Therefore, using Cauchy–Schwarz inequality again,

$$\begin{aligned}&\mathbb{E}\left[\left\|\int_{t_k}^{t_{k+1}}\Sigma_\epsilon^2\left(\tilde{\mathfrak{s}}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)-\tilde{\mathfrak{s}}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)\right)\mathrm{d}t\right\|^2\right]^{1/2}\\&\qquad\leq\sqrt{h}\left(\int_{t_k}^{t_{k+1}}\mathbb{E}\left[\left\|\Sigma_\epsilon^2\left(\tilde{\mathfrak{s}}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)-\tilde{\mathfrak{s}}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)\right)\right\|^2\right]\mathrm{d}t\right)^{1/2}.\end{aligned}$$

Then, using Lemma D.8,

$$\begin{aligned}\mathbb{E}\left[\left\|\Sigma_\epsilon^2\left(\tilde{\mathfrak{s}}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)-\tilde{\mathfrak{s}}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)\right)\right\|^2\right]&\leq\|\Sigma_\epsilon\|^2\,\mathbb{E}\left[\left\|\nabla\log\tilde{p}_{T-t}\left(\overrightarrow{\mathbf{U}}_t\right)-\nabla\log\tilde{p}_{T-t_k}\left(\overrightarrow{\mathbf{U}}_{t_k}\right)\right\|^2\right]\\&\leq C\,\|\Sigma_\epsilon\|^2\left(g(t_{k+1})-g(t_k)\right),\end{aligned}\tag{41}$$

with the function g defined in (59). This yields

$$\begin{aligned}\mathbb{E}\left[A_{4,k}\right]&\leq Ch\,\|\Sigma_\epsilon\|\sqrt{g(t_{k+1})-g(t_k)}\times\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right]^{1/2}\\&\leq h\frac{a-\eta}{6}\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_k}-\bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right]+h\frac{C\,\|\Sigma_\epsilon\|^2}{a-\eta}(g(t_{k+1})-g(t_k)),\end{aligned}$$

where we have used Young's inequality in the last inequality.

Bound of $\mathbb{E}[A_{5,k}]$. Using Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}\left[A_{5,k}\right]&=\int_{t_k}^{t_{k+1}}\mathbb{E}\left[\left(\overleftarrow{\mathbf{U}}_t-\overleftarrow{\mathbf{U}}_{t_k}\right)^\top\mathfrak{M}\tilde{A}_\epsilon\left(\overleftarrow{\mathbf{U}}_t-\overleftarrow{\mathbf{U}}_{t_k}\right)\right]\mathrm{d}t\\&\leq C\left\|\tilde{A}_\epsilon\right\|\mathbb{E}\left[\int_{t_k}^{t_{k+1}}\left\|\overleftarrow{\mathbf{U}}_s-\overleftarrow{\mathbf{U}}_{t_k}\right\|^2\mathrm{d}s\right].\end{aligned}$$

Then using Lemma D.3 as in (39),

$$\mathbb{E}\left[A_{5,k}\right]\leq Ch^2\left\|\tilde{A}_\epsilon\right\|\left(h\|A\|^2B^2+\|\Sigma_\epsilon\|^2 d\right),$$

with B defined as in 56.

Bound of $\mathbb{E}[A_{6,k}]$. Using Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}\left[A_{6,k}\right]&=\int_{t_k}^{t_{k+1}}\mathbb{E}\left[\left(\overleftarrow{\mathbf{U}}_t-\overleftarrow{\mathbf{U}}_{t_k}\right)^\top\mathfrak{M}\Sigma_\epsilon^2\left(\tilde{\mathfrak{s}}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)-\tilde{\mathfrak{s}}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)\right)\right]\mathrm{d}t\\&\leq C\int_{t_k}^{t_{k+1}}\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_t-\overleftarrow{\mathbf{U}}_{t_k}\right\|^2\right]^{1/2}\mathbb{E}\left[\left\|\Sigma_\epsilon^2\left(\tilde{\mathfrak{s}}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)-\tilde{\mathfrak{s}}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)\right)\right\|^2\right]^{1/2}\mathrm{d}t.\end{aligned}$$

Controlling the first term as in (39) and the second term as in (41), using Lemma D.8, together with Young's inequality, yields,

$$\mathbb{E}\left[A_{6,k}\right] \leq Ch^2(a-\eta)\left(h\|A\|^2B^2 + \|\Sigma_\epsilon\|^2 d\right) + Ch\frac{\|\Sigma_\epsilon\|^2\left(g(t_{k+1}) - g(t_k)\right)}{a-\eta}\,.$$

Final bound. Combining the upper bounds for $A_{1,k}$, $A_{2,k}$, $A_{3,k}$, $A_{4,k}$, $A_{5,k}$ and $A_{6,k}$, there exists a constant $C > 0$ such that

$$\begin{aligned}\mathbb{E}\left[\left\|\overleftarrow{\bar{\mathbf{U}}}_{t_{k+1}} - \bar{\mathbf{U}}_{t_{k+1}}\right\|_{\mathfrak{M}}^2\right] \leq \delta_k\mathbb{E}\left[\left\|\overleftarrow{\bar{\mathbf{U}}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right\|_{\mathfrak{M}}^2\right] + Ch\frac{\|\Sigma_\epsilon\|^2}{a-\eta}(g(t_{k+1}) - g(t_k))\\ + Ch^2\left(h\|A\|^2B^2 + \|\Sigma_\epsilon\|^2 d\right)\left((a-\eta) + \frac{a-\eta+2}{a-\eta}\|\tilde{A}_\epsilon\| \vee \|\tilde{A}_\epsilon\|^2\right)\\ + Ch^2\tilde{L}_{T-t_k}\left(h\|A\|^2B^2 + \|\Sigma_\epsilon\|^2 d\right)\,,\end{aligned}$$

with $\delta_k := 1 + h(C\tilde{L}_{T-t_k} - (a-\eta)/2)$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\left\|\overleftarrow{\bar{\mathbf{U}}}_{t_N} - \bar{\mathbf{U}}_{t_N}\right\|_{\mathfrak{M}}^2\right] \leq \left(\prod_{k=0}^{N-1}\delta_k\right)\mathbb{E}\left[\left\|\overleftarrow{\bar{\mathbf{U}}}_0 - \bar{\mathbf{U}}_0\right\|_{\mathfrak{M}}^2\right]\\ + Ch^2\left(h\|A\|^2B^2 + \|\Sigma_\epsilon\|^2 d\right)\left((a-\eta) + \frac{a-\eta+2}{a-\eta}\|\tilde{A}_\epsilon\| \vee \|\tilde{A}_\epsilon\|^2\right)\sum_{k=0}^{N-1}\prod_{j=k+1}^{N-1}\delta_j\\ + Ch\frac{\|\Sigma_\epsilon\|^2}{a-\eta}\sum_{k=0}^{N-1}(g(t_{k+1}) - g(t_k))\prod_{j=k+1}^{N-1}\delta_j\\ + Ch^2\left(h\|A\|^2B^2 + \|\Sigma_\epsilon\|^2 d\right)\sum_{k=0}^{N-1}\tilde{L}_{T-t_k}\prod_{j=k+1}^{N-1}\delta_j\,.\end{aligned} \tag{42}$$

First recall that the two processes share the same initialization, *i.e.* $\overleftarrow{\bar{\mathbf{U}}}_0 = \bar{\mathbf{U}}_0$. Note that, since $\exp(x) \geq 1 + x$, for $x \in \mathbb{R}$,

$$\begin{aligned}\prod_{j=k+1}^{N-1}\delta_j &\leq \exp\left(\sum_{j=k+1}^{N-1} h\left(C\tilde{L}_{T-t_j} - \frac{a-\eta}{2}\right)\right)\\ &\leq \exp\left(-\frac{a-\eta}{2}h(N-k-1) + C\sum_{j=k+1}^{N-1}h\tilde{L}_{T-t_j}\right)\\ &\leq \exp\left(-\frac{a-\eta}{2}h(N-k-1) + C\int_0^\infty \tilde{L}_s\mathrm{d}s\right)\\ &\leq \exp\left(-\frac{a-\eta}{2}h(N-k-1) + Ca^{-1}\right)\,,\end{aligned}$$

where we use the bound (26) from Proposition B.1. Combining this bound with the fact that $h \leq (1-\mathrm{e}^{-h})\mathrm{e}^h$, we obtain

$$h\exp\left(-\frac{a-\eta}{2}(N-k)h\right) \leq \frac{2\mathrm{e}^h}{a-\eta}\left(\exp\left(-\frac{a-\eta}{2}(N-k)h\right) - \exp\left(-\frac{a-\eta}{2}(N-k+1)h\right)\right)\,,$$

which then implies that

$$\begin{aligned}h\sum_{k=0}^{N-1}\prod_{j=k+1}^{N-1}\delta_j &\leq h\sum_{k=0}^{N-1}\mathrm{e}^{-\frac{a-\eta}{2}(N-k-2)h} \times \mathrm{e}^{Ca^{-1}}\\ &\leq \frac{2\mathrm{e}^h}{a-\eta}\sum_{k=0}^{N-1}\left(\mathrm{e}^{-\frac{a-\eta}{2}(N-k-2)h} - \mathrm{e}^{-\frac{a-\eta}{2}(N-k-1)h}\right) \times \mathrm{e}^{Ca^{-1}} \leq \frac{C\mathrm{e}^{Ca^{-1}}}{a-\eta}\,,\end{aligned}$$

increasing the value of the constant C if necessary. For the term involving $\tilde{L}_{T-t_k}$, note that

$$\begin{aligned}
&h\tilde{L}_{T-t_k}\exp(-\frac{a-\eta}{2}(N-k)h)\\
&\leq h\frac{C}{\sqrt{(N-k)h}}\exp(-\frac{a-\eta}{2}(N-k)h)\\
&\leq \frac{2\mathrm{e}^h}{a-\eta}\left(\Gamma\left(\frac{1}{2},\frac{a-\eta}{2}(N-k)h\right)-\Gamma\left(\frac{1}{2},\frac{a-\eta}{2}(N-k+1)h\right)\right),
\end{aligned}$$

where $\Gamma(a,b)$ denotes the Gamma function. Consequently,

$$\begin{aligned}
&h\sum_{k=0}^{N-1}\tilde{L}_{T-t_k}\prod_{j=k+1}^{N-1}\delta_j\\
&\leq \frac{2\mathrm{e}^h}{a-\eta}\sum_{k=0}^{N-1}\left(\Gamma\left(\frac{1}{2},\frac{a-\eta}{2}(N-k-2)h\right)-\Gamma\left(\frac{1}{2},\frac{a-\eta}{2}(N-k-1)h\right)\right)\times\mathrm{e}^{Ca^{-1}}\\
&\leq \frac{C\mathrm{e}^{Ca^{-1}}}{a-\eta}.
\end{aligned}$$

Moreover, we have that

$$\begin{aligned}
\sum_{k=0}^{N-1}(g(t_{k+1})-g(t_k))\prod_{j=k+1}^{N-1}\delta_j &\leq \mathrm{e}^{Ca^{-1}}\sum_{k=0}^{N-1}(g(t_{k+1})-g(t_k))\leq \mathrm{e}^{Ca^{-1}}g(t_N)\\
&\leq C\mathrm{e}^{Ca^{-1}}\mathbb{E}\left[\left\|\tilde{\mathfrak{s}}_\Delta\left(\overrightarrow{\mathbf{U}}_\Delta\right)\right\|^2\right].
\end{aligned}$$

Note that $\mathbb{E}[\|\tilde{\mathfrak{s}}_\Delta(\overrightarrow{\mathbf{U}}_\Delta)\|^2]$ corresponds to the relative Fisher information between p_Δ and π_∞. We can conclude for $\Delta\to 0$, following the argument of Conforti et al. (Lemma 3.9, 2025) and using Assumption H1, that $\mathcal{I}(\pi_{\mathrm{data}}\otimes\pi_v|\pi_\infty)=\mathbb{E}[\|\tilde{\mathfrak{s}}_0(\overrightarrow{\mathbf{U}}_0)\|^2]<\infty$. Then, applying (42) directly yields

$$\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{t_N}-\bar{\mathbf{U}}_{t_N}\right\|_{\mathfrak{M}}^2\right]\leq h\times C\left(h\|A\|^2B^2+\|\Sigma_\epsilon\|^2\left(d+\mathcal{I}(\pi_{\mathrm{data}}\otimes\pi_v|\pi_\infty)\right)\right)\frac{\mathrm{e}^{Ca^{-1}}}{a-\eta},$$

which concludes the proof.

□

Lemma B.3 (Approximation error)**.** *Assume that Assumptions H2 and H3 hold. Then, for any $\eta>0$, there exists a constant $C>0$ such that*

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{\mathbf{U}}_T^\infty\right),\mathcal{L}\left(\bar{\mathbf{U}}_T^\theta\right)\right)\leq C\frac{\|\Sigma_\epsilon\|^2}{a-\eta}M. \tag{43}$$

Proof. As in the proof of Lemma B.2, consider the synchronous coupling of the two processes $\bar{\mathbf{U}}^\infty$ and $\bar{\mathbf{U}}^\theta$ with the same initial condition $\bar{\mathbf{U}}_0^\infty=\bar{\mathbf{U}}_0^\theta$. We have

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{\mathbf{U}}_T^\infty\right),\mathcal{L}\left(\bar{\mathbf{U}}_T^\theta\right)\right)\leq\left\|\bar{\mathbf{U}}_T^\infty-\bar{\mathbf{U}}_T^\theta\right\|_{L_2}.$$

Fix $\Delta\geq 0$ such that $t_N=T-\Delta$ and note that for all $0\leq k\leq N-1$, from (12) and (13), we get

$$\begin{aligned}
&\bar{\mathbf{U}}_{t_{k+1}}^\infty-\bar{\mathbf{U}}_{t_{k+1}}^\theta\\
&=\bar{\mathbf{U}}_{t_k}^\infty-\bar{\mathbf{U}}_{t_k}^\theta+\int_{t_k}^{t_{k+1}}\left\{\tilde{A}_\epsilon\left(\bar{\mathbf{U}}_{t_k}^\infty-\bar{\mathbf{U}}_{t_k}^\theta\right)+\Sigma_\epsilon^2\left(\tilde{\mathfrak{s}}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}^\infty\right)-\tilde{s}_\theta\left(T-t_k,\bar{\mathbf{U}}_{t_k}^\theta\right)\right)\right\}\mathrm{d}t.
\end{aligned}$$

Taking $\mathfrak{M}$ as in the proof of Lemma B.2, we have

$$\left\|\bar{\mathbf{U}}_{t_{k+1}}^\infty-\bar{\mathbf{U}}_{t_{k+1}}^\theta\right\|_{\mathfrak{M}}^2=\left\|\bar{\mathbf{U}}_{t_k}^\infty-\bar{\mathbf{U}}_{t_k}^\theta\right\|_{\mathfrak{M}}^2+2B_{1,k}+2B_{2,k},$$

with

$$B_{1,k} = h\left(\bar{\mathbf{U}}^{\infty}_{t_k} - \bar{\mathbf{U}}^{\theta}_{t_k}\right)^{\top} \mathfrak{M}\tilde{A}_{\epsilon}\left(\bar{\mathbf{U}}^{\infty}_{t_k} - \bar{\mathbf{U}}^{\theta}_{t_k}\right) ,$$
$$B_{2,k} = h\left(\bar{\mathbf{U}}^{\infty}_{t_k} - \bar{\mathbf{U}}^{\theta}_{t_k}\right)^{\top} \mathfrak{M}\Sigma^2_{\epsilon}\left(\tilde{\mathfrak{s}}_{T-t_k}\left(\bar{\mathbf{U}}^{\infty}_{t_k}\right) - \tilde{s}_{\theta}\left(T-t_k, \bar{\mathbf{U}}^{\theta}_{t_k}\right)\right) .$$

Bound of $\mathbb{E}[B_{1,k}]$. From (38), we note that

$$\mathbb{E}\left[B_{1,k}\right] \leq -h\left(a-\eta\right)\mathbb{E}\left[\left\|\bar{\mathbf{U}}^{\infty}_{t_k} - \bar{\mathbf{U}}^{\theta}_{t_k}\right\|^2_{\mathfrak{M}}\right] .$$

Bound of $\mathbb{E}[B_{2,k}]$. We decompose the second term into the score and approximation components:

$$\begin{aligned}
&\mathbb{E}\left[B_{2,k}\right] \\
&= h\mathbb{E}\left[\left(\bar{\mathbf{U}}^{\infty}_{t_k} - \bar{\mathbf{U}}^{\theta}_{t_k}\right)^{\top} \mathfrak{M}\Sigma^2_{\epsilon}\left(\tilde{\mathfrak{s}}_{T-t_k}\left(\bar{\mathbf{U}}^{\infty}_{t_k}\right) - \tilde{\mathfrak{s}}_{T-t_k}\left(\bar{\mathbf{U}}^{\theta}_{t_k}\right)\right)\right] \\
&\quad + h\mathbb{E}\left[\left(\bar{\mathbf{U}}^{\infty}_{t_k} - \bar{\mathbf{U}}^{\theta}_{t_k}\right)^{\top} \mathfrak{M}\Sigma^2_{\epsilon}\left(\tilde{\mathfrak{s}}_{T-t_k}\left(\bar{\mathbf{U}}^{\theta}_{t_k}\right) - \tilde{s}_{\theta}\left(T-t_k, \bar{\mathbf{U}}^{\theta}_{t_k}\right)\right)\right] \\
&\leq hC\tilde{L}_{T-t_k}\mathbb{E}\left[\left\|\bar{\mathbf{U}}^{\infty}_{t_k} - \bar{\mathbf{U}}^{\theta}_{t_k}\right\|^2_{\mathfrak{M}}\right] + h\left\|\Sigma_{\epsilon}\right\| M\mathbb{E}\left[\left\|\bar{\mathbf{U}}^{\infty}_{t_k} - \bar{\mathbf{U}}^{\theta}_{t_k}\right\|_{\mathfrak{M}}\right] \\
&\leq hC\tilde{L}_{T-t_k}\mathbb{E}\left[\left\|\bar{\mathbf{U}}^{\infty}_{t_k} - \bar{\mathbf{U}}^{\theta}_{t_k}\right\|^2_{\mathfrak{M}}\right] + h\frac{a-\eta}{2}\mathbb{E}\left[\left\|\bar{\mathbf{U}}^{\infty}_{t_k} - \bar{\mathbf{U}}^{\theta}_{t_k}\right\|^2_{\mathfrak{M}}\right] + hC\left\|\Sigma_{\epsilon}\right\|^4 M^2 ,
\end{aligned}$$

where we have used Young's inequality in the last inequality, for $C > 0$ a universal constant (which may change from line to line) depending only on the eigenvalues of the matrix $\mathfrak{M}$ or constant factors.
Final bound. Combining the bounds on $B_{1,k}$ and $B_{2,k}$, there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\left\|\bar{\mathbf{U}}^{\infty}_{t_{k+1}} - \bar{\mathbf{U}}^{\theta}_{t_{k+1}}\right\|^2_{\mathfrak{M}}\right] \leq \delta_k\mathbb{E}\left[\left\|\bar{\mathbf{U}}^{\infty}_{t_k} - \bar{\mathbf{U}}^{\theta}_{t_k}\right\|^2_{\mathfrak{M}}\right] + h\frac{C}{a-\eta}\left\|\Sigma_{\epsilon}\right\|^4 M^2 ,$$

with $\delta_k := 1 + h\left(C\tilde{L}_{T-t_k} - (a-\eta)/2\right)$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{\mathbf{U}}^{\infty}_{t_N} - \bar{\mathbf{U}}^{\theta}_{t_N}\right\|^2_{\mathfrak{M}}\right] \\
&\leq \prod_{j=k+1}^{N-1}\delta_j\mathbb{E}\left[\left\|\bar{\mathbf{U}}^{\infty}_0 - \bar{\mathbf{U}}^{\theta}_0\right\|^2_{\mathfrak{M}}\right] + h\frac{C}{a-\eta}\left\|\Sigma_{\epsilon}\right\|^4 M^2\sum_{k=0}^{N-1}\prod_{j=k+1}^{N-1}\delta_j \\
&\leq h\frac{C}{a-\eta}\left\|\Sigma_{\epsilon}\right\|^4 M^2\sum_{k=0}^{N-1}\prod_{j=k+1}^{N-1}\delta_j ,
\end{aligned}$$

where we used that $\bar{\mathbf{U}}^{\infty}_0 = \bar{\mathbf{U}}^{\theta}_0$. Following the same argument as in Lemma B.2 (discretization error term), we obtain

$$\mathbb{E}\left[\left\|\bar{\mathbf{U}}^{\infty}_{t_N} - \bar{\mathbf{U}}^{\theta}_{t_N}\right\|^2_{\mathfrak{M}}\right] \leq C\frac{\left\|\Sigma_{\epsilon}\right\|^4 M^2}{(a-\eta)^2} .$$

We conclude the proof by taking the limit as $\Delta \to 0$ together with Fatou's lemma. □

Lemma B.4 (Mixing time)**.** *Assume that H2 holds. Then, for all $\eta > 0$, there exists a constant $C > 0$ such that*

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{\mathbf{U}}_T\right), \mathcal{L}\left(\bar{\mathbf{U}}^{\infty}_T\right)\right) \leq C\mathrm{e}^{Ca^{-1}} \times Te^{-\frac{3}{2}(a-\eta)T}\mathcal{W}_2\left(\pi_{\mathrm{data}}\otimes\pi_v, \pi_{\infty}\right) . \tag{44}$$

Proof. Consider a synchronous coupling of the continuous-time interpolations $(\bar{\mathbf{U}}_t)_{t\in[0,T]}$ and $(\bar{\mathbf{U}}^{\infty}_t)_{t\in[0,T]}$, defined in (12), with initialization

$$\mathcal{W}_2\left(\pi_{\infty}, \mathcal{L}\left(\overrightarrow{\bar{\mathbf{U}}}_T\right)\right) = \left\|\bar{\mathbf{U}}_0 - \bar{\mathbf{U}}^{\infty}_0\right\|_{L_2} .$$

By definition of the $\mathcal{W}_2$ distance,

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{\mathbf{U}}_T\right), \mathcal{L}\left(\bar{\mathbf{U}}^{\infty}_T\right)\right) \leq \left\|\bar{\mathbf{U}}_T - \bar{\mathbf{U}}^{\infty}_T\right\|_{L_2} .$$

Analogously to the proof of Lemma B.2 and Lemma B.3, fix $\Delta \geq 0$ such that $t_N = T - \delta$, and note that for $t \in [t_k, t_{k+1}]$, we have that

$$\left\| \bar{\mathbf{U}}_{t_{k+1}} - \bar{\mathbf{U}}^{\infty}_{t_{k+1}} \right\|^2_{\mathfrak{M}} = \left\| \bar{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}^{\infty}_{t_k} \right\|^2_{\mathfrak{M}} + C_k \,,$$

with

$$C_k = h2 \left(\bar{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}^{\infty}_{t_k} \right)^{\top} \mathfrak{M} \left\{ \tilde{A}_{\epsilon} \left(\bar{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}^{\infty}_{t_k} \right) + \Sigma^2_{\epsilon} \left(\tilde{\mathfrak{s}}_{T-t_k} \left(\bar{\mathbf{U}}_{t_k} \right) - \tilde{\mathfrak{s}}_{T-t_k} \left(\bar{\mathbf{U}}^{\infty}_{t_k} \right) \right) \right\} .$$

Similarly to Lemma B.2, we use that, for any fixed $\eta > 0$, we have

$$\mathfrak{M} \tilde{A}_{\epsilon} \preccurlyeq -(a - \eta) \mathfrak{M} \,.$$

Therefore,

$$\left(\bar{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}^{\infty}_{t_k} \right)^{\top} \mathfrak{M} \tilde{A}_{\epsilon} \left(\bar{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}^{\infty}_{t_k} \right) \leq -(a - \eta) \left\| \bar{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}^{\infty}_{t_k} \right\|^2_{\mathfrak{M}} \,,$$

and using Proposition B.1 there exists $C > 0$ a universal constant (depending only on the eigenvalues of the matrix $\mathfrak{M}$ or constant factors) such that

$$\left(\bar{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}^{\infty}_{t_k} \right)^{\top} \mathfrak{M} \Sigma^2 \left(\tilde{\mathfrak{s}}_{T-t_k} \left(\bar{\mathbf{U}}_{t_k} \right) - \tilde{\mathfrak{s}}_{T-t_k} \left(\bar{\mathbf{U}}^{\infty}_{t_k} \right) \right) \leq C \tilde{L}_{T-t_k} \left\| \bar{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}^{\infty}_{t_k} \right\|^2_{\mathfrak{M}} \,,$$

it follows that

$$\mathbb{E} \left[C_k \right] \leq h \left(C \tilde{L}_{T-t_k} - (a - \eta) \right) \mathbb{E} \left[\left\| \bar{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}^{\infty}_{t_k} \right\|^2_{\mathfrak{M}} \right] .$$

As a consequence,

$$\mathbb{E} \left[\left\| \bar{\mathbf{U}}_{t_N} - \bar{\mathbf{U}}^{\infty}_{t_N} \right\|^2_{\mathfrak{M}} \right] \leq \mathbb{E} \left[\left\| \bar{\mathbf{U}}_0 - \bar{\mathbf{U}}^{\infty}_0 \right\|^2_{\mathfrak{M}} \right] \prod_{\ell=0}^{N-1} \delta'_{\ell} \,,$$

with $\delta'_{\ell} = 1 + h(C \tilde{L}_{T-t_k} - (a - \eta))$. Since $\exp(x) \geq 1 + x$, for $x \in \mathbb{R}$, we have that

$$\begin{aligned} \prod_{\ell=0}^{N-1} \delta'_{\ell} &\leq \mathrm{e}^{\sum_{k=0}^{N-1} h \left(C \tilde{L}_{T-t_k} - (a-\eta) \right)} \\ &\leq \mathrm{e}^{-(a-\eta)T + C \sum_{k=0}^{N-1} h \tilde{L}_{T-t_k}} \\ &\leq \mathrm{e}^{-(a-\eta)T + C \int_0^{\infty} \tilde{L}_s \mathrm{d}s} \\ &\leq \mathrm{e}^{-(a-\eta)T + C a^{-1}} \,, \end{aligned}$$

thus,

$$\mathbb{E} \left[\left\| \bar{\mathbf{U}}_{t_N} - \bar{\mathbf{U}}^{\infty}_{t_N} \right\|^2_{\mathfrak{M}} \right] \leq \mathrm{e}^{C a^{-1}} \mathrm{e}^{-(a-\eta)T} \mathbb{E} \left[\left\| \bar{\mathbf{U}}_0 - \bar{\mathbf{U}}^{\infty}_0 \right\|^2_{\mathfrak{M}} \right] ,$$

which implies, taking the limit as $\Delta \to 0$ together with Fatou's lemma, that

$$\mathcal{W}^2_2 \left(\mathcal{L} \left(\bar{\mathbf{U}}_T \right), \mathcal{L} \left(\bar{\mathbf{U}}^{\infty}_T \right) \right) \leq C \mathrm{e}^{C a^{-1}} \mathrm{e}^{-(a-\eta)T} \left\| \bar{\mathbf{U}}_0 - \bar{\mathbf{U}}^{\infty}_0 \right\|^2_{L_2} .$$

Moreover, similarly to the backward, the forward process also satisfies the following contraction property (Lemma A.4),

$$\left\| \bar{\mathbf{U}}_0 - \bar{\mathbf{U}}^{\infty}_0 \right\|_{L_2} = \mathcal{W}_2 \left(\pi_{\infty}, \mathcal{L} \left(\overrightarrow{\bar{\mathbf{U}}}_T \right) \right) \leq C T \mathrm{e}^{-(a-\eta)T} \mathcal{W}_2 \left(\pi_{\mathrm{data}} \otimes \pi_v, \pi_{\infty} \right) ,$$

yielding (44). □

C Proof of Theorem 3.2

In this section, we prove Theorem 3.2. To establish this result, we work with the (unmodified) score function rather than the modified one used previously. Similarly to the previous section, we introduce

the continuous time interpolation $(\bar{\mathbf{U}}_t)_{t\in[0,T]}$ of the Euler scheme for the time-reversed process $(\overleftarrow{\mathbf{U}}_t)_{t\in[0,T]}$ defined as the Itô process, for $t \in [t_k, t_{k+1}]$,

$$\bar{\mathbf{U}}_t = \bar{\mathbf{U}}_{t_k} + \left(-A\bar{\mathbf{U}}_{t_k} + \Sigma_\varepsilon^2 \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)(t - t_k) + \Sigma_\varepsilon\left(B_t - B_{t_k}\right), \tag{45}$$

when initialized at p_T (*i.e.*, $\bar{\mathbf{U}}_0 \sim p_T$). When initialized at π_∞, we write $(\bar{\mathbf{U}}_t^\infty)_{t\in[0,T]}$ this Itô process. We also introduce the continuous time Euler scheme $(\bar{\mathbf{U}}_t^\theta)_{t\in[0,T]}$ in which the true, unknown score function is replaced by a neural network approximation s_θ, and defined for $t \in [t_k, t_{k+1}]$ as

$$\bar{\mathbf{U}}_t^\theta = \bar{\mathbf{U}}_{t_k}^\theta + \left(-A\bar{\mathbf{U}}_{t_k}^\theta + \Sigma_\varepsilon^2 s_\theta\left(t_k, \bar{\mathbf{U}}_{t_k}^\theta\right)\right)(t - t_k) + \Sigma_\varepsilon\left(B_t - B_{t_k}\right), \tag{46}$$

where $\bar{\mathbf{U}}_0^\theta \sim \pi_\infty$.

We first establish the propagation of regularity properties: strong log-concavity propagation (Proposition C.1) and Lipschitz regularity propagation (Proposition C.2), followed by the proof of Theorem 3.2. To this end, we decompose the generation error into the sum of the discretization error (Lemma C.3), the approximation error (Lemma C.4), and the mixing time error (Lemma C.5), as in Theorem 3.1.

C.1 Propagation of the regularity assumptions

Proposition C.1. *Assume that H2′ holds. Then for all $t \in [0, T]$ and all $u \in \mathbb{R}^{2d}$,*

$$\nabla^2 \log p_t(u) \preccurlyeq -\alpha_t \mathbf{I}_{2d}\,,$$

where

$$\alpha_t = \left(\frac{1}{(\alpha_0 \wedge v^{-2})\sigma_{\min}^2(e^{-tA})} + \lambda_{\max}(\Sigma_{0,t})\right)^{-1}. \tag{47}$$

Proof. Similar to Proposition B.1 recall the following equality in law given by the modified kinetic OU process (9)

$$\overrightarrow{\mathbf{U}}_t \overset{\mathcal{L}}{=} \mathrm{e}^{tA}\overrightarrow{\mathbf{U}}_0 + \sqrt{\Sigma_{0,t}}G\,,$$

with $\overrightarrow{\mathbf{U}}_0 \sim \pi_{\mathrm{data}} \otimes \pi_v$, $G \sim \mathcal{N}(0, \mathbf{I}_{2d})$, where G and $\overrightarrow{\mathbf{U}}_0$ are independent, and $\Sigma_{0,t}$ is defined in (20). Writing $q_{t|0}$ the conditional density of $\overrightarrow{\mathbf{U}}_t$ given $\overrightarrow{\mathbf{U}}_0$, we have

$$p_t(u_t) = \det\left(\mathrm{e}^{-tA}\right)\int_{\mathbb{R}^{2d}} p_0\left(\mathrm{e}^{-tA}z\right)\det\left(2\pi\Sigma_{0,t}\right)^{-1/2}\exp\left(-\frac{1}{2}\Big(u_t - z\Big)^\top \Sigma_{0,t}^{-1}\Big(u_t - z\Big)\right)\mathrm{d}z.$$

Since π_{data} is α_0–strongly log-concave and π_v is a centered Gaussian with covariance $v^2\mathbf{I}_d$, their product (*i.e.* the probability density function of $\overrightarrow{\mathbf{U}}_0$) satisfies

$$\nabla^2 \log p_0(z) \preccurlyeq -(\alpha_0 \wedge v^{-2})\,\mathbf{I}_{2d}\,.$$

Consequently, for any $z \in \mathbb{R}^{2d}$,

$$\nabla^2 \log p_0(\mathrm{e}^{-tA}z) \preccurlyeq -(\alpha_0 \wedge v^{-2})(\mathrm{e}^{-tA})^\top \mathrm{e}^{-tA}\,.$$

Finally, using Saumard and Wellner (2014), p_t is strongly log-concave with constant

$$\alpha_t = \left(\frac{1}{(\alpha_0 \wedge v^{-2})\sigma_{\min}^2(e^{-tA})} + \lambda_{\max}(\Sigma_{0,t})\right)^{-1}.$$

□

Proposition C.2. *Assume that H2′ holds. Then, for all $t > 0$, $\nabla \log p_t$ is L_t-Lipschitz: for all $u \in \mathbb{R}^{2d}$,*

$$\left\|\nabla^2 \log p_t(u)\right\| \le L_t \le \min\left\{\mathfrak{h}_{1,t}; \mathfrak{h}_{2,t}\right\}.$$

where

$$\mathfrak{h}_{1,t} = \left(1 + (a+1)^2 t\right)^2 \mathrm{e}^{2ta}\max\left\{L_0, v^{-2}\right\}$$
$$\mathfrak{h}_{2,t} = \frac{4}{\lfloor \sigma^2 \min\{a, 1/a\} - (\sigma^2 \max\{a, 1/a\} + 5\varepsilon^2 a^{-1})\,\mathrm{e}^{-2at}\rfloor_+}.$$

Proof. Following "*Step 1: Lower bound on* $\nabla^2 \log p_t$" in the proof of Proposition B.1, we obtain for all $t > 0$

$$\nabla^2 \log p_t(u) \succcurlyeq -\min\left\{\mathfrak{h}_{1,t}; \mathfrak{h}_{2,t}\right\} \mathbf{I}_{2d}\,,$$

where

$$\mathfrak{h}_{1,t} = \left(1 + (a+1)^2 t\right)^2 \mathrm{e}^{2ta} \max\left\{L_0, v^{-2}\right\}$$
$$\mathfrak{h}_{2,t} = \frac{4}{\lfloor \sigma^2 \min\{a, 1/a\} - (\sigma^2 \max\{a, 1/a\} + 5\varepsilon^2 a^{-1})\,\mathrm{e}^{-2at} \rfloor_+}\,.$$

Moreover, Proposition C.1 implies that

$$\begin{aligned}\nabla^2 \log p_t(u) &\preccurlyeq -\alpha_t \mathbf{I}_{2d} \\ &\preccurlyeq 0_{2d\times 2d}\,,\end{aligned}$$

where α_t defined as in (47). Consequently,

$$\left\|\nabla^2 \log p_t(u)\right\| \leq \min\left\{\mathfrak{h}_{1,t}; \mathfrak{h}_{2,t}\right\}\,.$$

□

C.2 Proofs of the main results

Lemma C.3 (Discretization error)**.** *Assume that H2′ holds and let* $\varepsilon > 0$*. If the step size* h *satisfies*

$$0 < h < \frac{2\min_k \alpha_{t_k}\left(\sigma^2 \wedge \varepsilon^2\right) - (\sigma-\varepsilon)^2 \max_k L_{t_k} - (a+1)^2}{\|A\|^2 + (\varepsilon^4 + \sigma^4)\max_k L_{t_k}^2 + 2\left(\sigma^2 \vee \varepsilon^2\right)\|A\| \max_k L_{t_k}}\,,$$

then, there exists $\delta_\varepsilon > 0$ *such that* $\mathcal{W}_2\left(\mathcal{L}\big(\overleftarrow{\mathbf{U}}_T\big), \mathcal{L}\big(\bar{\mathbf{U}}_T\big)\right) \leq 2\sqrt{h} C_a(\varepsilon)/\delta_\varepsilon$ *where*

$$C_a(\varepsilon) = \Big(2\|A\|^4 B_\varepsilon + 4d(a^2\sigma^2 + \varepsilon)^2 \Lambda_\varepsilon^*(T)\Big)h + 4d\Big(\|A\|^2 + \sigma^4 \sup_{t\in[0,T]} L_{T-t}^2\Big)\,, \tag{48}$$

with

$$\Lambda_\varepsilon^*(T) = \min\left\{\frac{2a\big(1+(a+1)^2 T\big)^2}{\min\{\varepsilon^2, \sigma^2\}}, \frac{4}{\sigma^2\min\{a,1/a\} - \big(\sigma^2\max\{a,1/a\} + 5a\varepsilon^{-2}\big)\mathrm{e}^{-2aT}}\right\},$$

such that $\sup_{T>0} \Lambda_\varepsilon^*(T) < +\infty$ *and*

$$B_\varepsilon := \max_{t\in[0,T]} \big(1+(a+1)^2(T-t)\big)^2 \mathrm{e}^{-2a(T-t)} \|\overrightarrow{\mathbf{U}}_0\|_{L^2}^2 + \frac{d}{2}\left(\sigma^2 \max\{a, 1/a\} + \frac{5\varepsilon^2}{a}\right)\,. \tag{49}$$

Proof. Consider a synchronous coupling for $(\overleftarrow{\mathbf{U}}_t)_{t\in[0,T]}$ and $(\bar{\mathbf{U}}_t)_{t\in[0,T]}$ *i.e.*, use the same Brownian motion to drive the two processes, with the same initial point, *i.e.*, $\overleftarrow{\mathbf{U}}_0 = \bar{\mathbf{U}}_0$. Then it holds, that

$$\mathcal{W}_2\left(\mathcal{L}(\overleftarrow{\mathbf{U}}_T), \mathcal{L}(\bar{\mathbf{U}}_T)\right) \leq \left\|\overleftarrow{\mathbf{U}}_T - \bar{\mathbf{U}}_T\right\|_{L_2}\,.$$

Fix $\Delta \geq 0$ such that $t_N = T - \Delta$ and note that for all $0 \leq k \leq N-1$,

$$\begin{aligned}&\left\|\overleftarrow{\mathbf{U}}_{t_{k+1}} - \bar{\mathbf{U}}_{t_{k+1}}\right\|_{L_2} \\ &\quad= \left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k} + \int_{t_k}^{t_{k+1}} \left\{-A\left(\overleftarrow{\mathbf{U}}_t - \bar{\mathbf{U}}_{t_k}\right) + \Sigma_\varepsilon^2\left(\mathsf{s}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right) - \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\right\} \mathrm{d}t\right\|_{L_2} \\ &\quad\leq A_{1,k} + A_{2,k}\,,\end{aligned}$$

where

$$A_{1,k} = \left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k} + \int_{t_k}^{t_{k+1}} \left\{-A\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right) + \Sigma_\varepsilon^2\left(\mathsf{s}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right) - \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\right\} \mathrm{d}t\right\|_{L_2}\,,$$

$$A_{2,k} = \left\|\int_{t_k}^{t_{k+1}} \left\{-A\left(\overleftarrow{\mathbf{U}}_t - \overleftarrow{\mathbf{U}}_{t_k}\right) + \Sigma_\varepsilon^2\left(\mathsf{s}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right) - \mathsf{s}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)\right)\right\} \mathrm{d}t\right\|_{L_2}\,.$$

For the first term, note that,

$$\begin{aligned}
A_{1,k}^2 &= \left\|(\mathbf{I}_{2d} - hA)\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right) + h\Sigma_\varepsilon^2\left(\mathsf{s}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right) - \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\right\|_{L_2}^2 \\
&= \left\|(\mathbf{I}_{2d} - hA)\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right)\right\|_{L_2}^2 + \left\|h\Sigma_\varepsilon^2\left(\mathsf{s}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right) - \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\right\|_{L_2}^2 \\
&\qquad + 2h\mathbb{E}\left[\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right)^\top (\mathbf{I}_{2d} - hA)^\top \Sigma_\varepsilon^2\left(\mathsf{s}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right) - \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\right] .
\end{aligned}$$

By Proposition C.2, it follows that the score at time t is L_t-Lipschitz continuous for all $t \in [0,T]$, in particular,

$$\left\|h\Sigma_\varepsilon^2\left(\mathsf{s}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right) - \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\right\|_{L_2}^2 \leq h^2(\varepsilon^4+\sigma^4)L_{T-t_k}^2\left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right\|_{L_2}^2 .$$

Therefore,

$$\begin{aligned}
A_{1,k}^2 &\leq \mathbb{E}\left[\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right)^\top \left((\mathbf{I}_{2d} - hA)^\top(\mathbf{I}_{2d} - hA) + h^2(\varepsilon^4+\sigma^4)L_{T-t_k}^2\mathbf{I}_{2d}\right)\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right)\right] \\
&\qquad + 2h\mathbb{E}\left[\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right)^\top (\mathbf{I}_{2d} - hA)^\top \Sigma_\varepsilon^2\left(\mathsf{s}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right) - \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right)\right)\right] .
\end{aligned}$$

For all $0 \leq t \leq T$, let $C_{t,k} := \int_0^1 \nabla^2 \log p_t\big(\overleftarrow{\mathbf{U}}_{t_k} - \gamma(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k})\big)\mathrm{d}\gamma$ and write $\mathbf{A}_h = \mathbf{I}_{2d} - hA$ so that,

$$\begin{aligned}
A_{1,k}^2 &\leq \mathbb{E}\left[\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right)^\top \left(\mathbf{A}_h^\top\mathbf{A}_h + h^2(\varepsilon^4+\sigma^4)L_{T-t_k}^2\mathbf{I}_{2d} + 2h\mathbf{A}_h^\top\Sigma_\varepsilon^2 C_{T-t_k,k}\right)\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right)\right] \\
&\leq \left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right\|_{L_2}^2 + h\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right)^\top M_h(\varepsilon)\left(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right) ,
\end{aligned}$$

where

$$M_h(\varepsilon) = -(A^\top + A) + 2\Sigma_\varepsilon^2 C_{T-t_k,k} + h\left(A^\top A + (\varepsilon^4+\sigma^4)L_{T-t_k}^2\mathbf{I}_{2d} - 2A^\top\Sigma_\varepsilon^2 C_{T-t_k,k}\right) .$$

In order to control $(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k})^\top M_h(\varepsilon)(\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k})$, it is enough to control the eigenvalues of $\tilde{M}_h(\varepsilon)$ where

$$\begin{aligned}
\tilde{M}_h(\varepsilon) &= \frac{1}{2}(M_h(\varepsilon) + M_h(\varepsilon)^\top) \\
&= -(A^\top + A) + (\Sigma_\varepsilon^2 C_{T-t_k,k} + C_{T-t_k,k}\Sigma_\varepsilon^2) \\
&\qquad + h\left\{A^\top A + (\varepsilon^4+\sigma^4)L_{T-t_k}^2\mathbf{I}_{2d} - \left(A^\top\Sigma_\varepsilon^2 C_{T-t_k,k} + C_{T-t_k,k}\Sigma_\varepsilon^2 A\right)\right\} .
\end{aligned}$$

Noting that,

$$(\Sigma_\varepsilon^2 C_{T-t_k,k} + C_{T-t_k,k}\Sigma_\varepsilon^2) = 2\Sigma_\varepsilon C_{T-t_k,k}\Sigma_\varepsilon + \Sigma_\varepsilon^2 C_{T-t_k,k} + C_{T-t_k,k}\Sigma_\varepsilon^2 - 2\Sigma_\varepsilon C_{T-t_k,k}\Sigma_\varepsilon$$

By Proposition C.1,

$$\begin{aligned}
\Sigma_\varepsilon C_{T-t_k,k}\Sigma_\varepsilon &\preccurlyeq -\alpha_{T-t_k}\lambda_{\min}\left(\Sigma_\varepsilon^2\right)\mathbf{I}_{2d} \\
&\preccurlyeq -\alpha_{T-t_k}\left(\sigma^2 \wedge \varepsilon^2\right)\mathbf{I}_{2d} ,
\end{aligned}$$

and simple calculations yields

$$\Sigma_\varepsilon^2 C_{T-t_k,k} + C_{T-t_k,k}\Sigma_\varepsilon^2 - 2\Sigma_\varepsilon C_{T-t_k,k}\Sigma_\varepsilon = (\sigma-\varepsilon)^2\begin{pmatrix} 0_{d\times d} & C_{T-t_k,k}^{12} \\ C_{T-t_k,k}^{12\,\top} & 0_{d\times d}\end{pmatrix} ,$$

where $C_{T-t_k,k}^{12}$ denotes the block anti diagonal elements of C_{T-t_k}. Hence,

$$\begin{aligned}
\Sigma_\varepsilon^2 C_{T-t_k,k} + C_{T-t_k,k}\Sigma_\varepsilon^2 - 2\Sigma_\varepsilon C_{T-t_k,k}\Sigma_\varepsilon &\preccurlyeq (\sigma-\varepsilon)^2\left\|C_{T-t_k,k}\right\|\mathbf{I}_{2d} \\
&\preccurlyeq (\sigma-\varepsilon)^2 L_{T-t_k}\mathbf{I}_{2d} ,
\end{aligned}$$

where we used Proposition C.2 in the last line. It follows that,

$$\tilde{M}_h(\varepsilon) \preccurlyeq -\lambda_{\min}(A^\top + A)\mathbf{I}_{2d} - 2\alpha_{T-t_k}\left(\sigma^2 \wedge \varepsilon^2\right)\mathbf{I}_{2d} + (\sigma-\varepsilon)^2 L_{T-t_k}\mathbf{I}_{2d} + h\left(\|A\|^2 + (\varepsilon^4+\sigma^4)L_{T-t_k}^2 + 2\left(\sigma^2 \vee \varepsilon^2\right)\|A\| L_{T-t_k}\right)\mathbf{I}_{2d}\,.$$

Therefore, using that $\lambda_{\min}(A^\top + A) = -(a+1)^2$, $\tilde{M}_h(\varepsilon)$ is negative when h is chosen so that

$$h < \frac{2\min_k \alpha_{t_k}\left(\sigma^2 \wedge \varepsilon^2\right) - (\sigma-\varepsilon)^2 \max_k L_{t_k} - (a+1)^2}{\|A\|^2 + (\varepsilon^4+\sigma^4)\max_k L_{t_k}^2 + 2\left(\sigma^2 \vee \varepsilon^2\right)\|A\| \max_k L_{t_k}}\,. \tag{50}$$

It follows that when h satisfies (50), there exists $\delta_\varepsilon > 0$ such that

$$A_{1,k} \le \sqrt{1-h\delta_\varepsilon}\left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right\|_{L_2}\,.$$

For the second term $A_{2,k}$, note the backward drift function as $b(t,u) = -Au + \Sigma_\varepsilon^2 \mathsf{s}_{T-t}(u)$ so that

$$A_{2,k}^2 = \mathbb{E}\left[\left\|\int_{t_k}^{t_{k+1}} \left(b(t,\overleftarrow{\mathbf{U}}_t) - b(t_k,\overleftarrow{\mathbf{U}}_{t_k})\right)\mathrm{d}t\right\|^2\right]\,.$$

Applying Lemma D.6 combining with Itô's formula, we obtain

$$\begin{aligned}
\mathrm{d}b(t,\overleftarrow{\mathbf{U}}_t) &= -A\mathrm{d}\overleftarrow{\mathbf{U}}_t + \Sigma_\varepsilon^2 \mathrm{d}\mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\\
&= \left\{AA\overleftarrow{\mathbf{U}}_t - A\Sigma_\varepsilon^2 \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t) + \Sigma_\varepsilon^2 A^T \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\right\}\mathrm{d}t + \left(A + \Sigma_\varepsilon^2 \nabla^2 \log p_{T-t}(\overleftarrow{\mathbf{U}}_t)\right)\Sigma_\varepsilon \mathrm{d}B_t\\
&= AA\overleftarrow{\mathbf{U}}_t \mathrm{d}t + \left(\Sigma_\varepsilon^2 A^\top - A\Sigma_\varepsilon^2\right)\mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\mathrm{d}t + \left(A + \Sigma_\varepsilon^2 \nabla^2 \log p_{T-t}(\overleftarrow{\mathbf{U}}_t)\right)\Sigma_\varepsilon \mathrm{d}B_t\,.
\end{aligned}$$

Using $H_s = AA\overleftarrow{\mathbf{U}}_s + \left(\Sigma_\varepsilon^2 A^\top - A\Sigma_\varepsilon^2\right)\mathsf{s}_{T-s}(\overleftarrow{\mathbf{U}}_s)$ and $K_s = \left(A + \Sigma_\varepsilon^2 \nabla^2 \log p_{T-s}(\overleftarrow{\mathbf{U}}_s)\right)\Sigma_\varepsilon$ we have that

$$\begin{aligned}
A_{2,k}^2 &= \mathbb{E}\left[\left\|\int_{t_k}^{t_{k+1}}\int_{t_k}^{t} H_s \mathrm{d}s\mathrm{d}t + \int_{t_k}^{t_{k+1}}\int_{t_k}^{t} K_s \mathrm{d}B_s \mathrm{d}t\right\|^2\right]\,,\\
&\le 2h\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\|\int_{t_k}^{t} H_s \mathrm{d}s\right\|^2 \mathrm{d}t\right] + 2h^2 \mathbb{E}\left[\sup_{t\in[t_k,t_{k+1}]}\left\|\int_{t_k}^{t} K_s \mathrm{d}B_s\right\|^2\right]\,,
\end{aligned}$$

by convexity of $\|\cdot\|^2$. Using again the convexity (or applying Cauchy-Schwartz inequality) we have $\mathbb{E}[\|\int_{t_k}^t H_s \mathrm{d}s\|^2] \le h\int_{t_k}^t \mathbb{E}[\|H_s\|^2]\mathrm{d}s$ and then

$$A_{2,k}^2 \le h^4 \sup_{t\in[t_k,t_{k+1}]} \mathbb{E}\left[\|H_t\|^2\right] + 2h^2 \mathbb{E}\left[\sup_{t\in[t_k,t_{k+1}]}\left\|\int_{t_k}^{t} K_s \mathrm{d}B_s\right\|^2\right]\,. \tag{51}$$

First we have for $t \in [0,T]$,

$$\mathbb{E}\left[\|H_t\|^2\right] \le 2\|A\|_2^4 \mathbb{E}\left[\|\overleftarrow{\mathbf{U}}_t\|^2\right] + 2\left\|\Sigma_\varepsilon^2 A^\top - A\Sigma_\varepsilon^2\right\|^2 \mathbb{E}\left[\left\|\mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\right\|^2\right]\,,$$

and by Lemma D.3 and Lemma D.5 we get

$$\mathbb{E}\left[\|H_t\|^2\right] \le 2\|A\|^4 B_\varepsilon + 2\left\|\Sigma_\varepsilon^2 A^\top - A\Sigma_\varepsilon^2\right\|^2 \frac{2d}{\lambda_{\min}(\Sigma_{0,T-t})}\,,$$

where B_ε is defined in (49). By Lemma A.3 we get

$$\max_{t\in[0,T]} \mathbb{E}\left[\|H_t\|^2\right] \le 2\|A\|^4 B_\varepsilon + 4d(a^2\sigma^2+\varepsilon)^2 \Lambda_\varepsilon^*(T)\,, \tag{52}$$

with

$$\Lambda_\varepsilon^*(T) = \min\left\{\frac{2a\left(1+(a+1)^2T\right)^2}{\min\{\varepsilon^2,\sigma^2\}}, \frac{4}{\sigma^2\min\{a,1/a\} - \left(\sigma^2\max\{a,1/a\} + 5a^{-1}\varepsilon^2\right)\mathrm{e}^{-2aT}}\right\}\,,$$

such that $\sup_{T>0} \Lambda^*_\varepsilon(T) < +\infty$.

Now by Doob's inequality and Itô's isometry, we have

$$\mathbb{E}\left[\sup_{t\in[t_k,t_{k+1}]}\left\|\int_{t_k}^t K_s \mathrm{d}B_s\right\|^2\right] = \int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\|A + \Sigma_\varepsilon^2 \nabla^2 \log p_{T-s}(\overleftarrow{\mathbf{U}}_s)\right\|_F^2\right] \mathrm{d}s\,,$$

where $\|\cdot\|_F$ is the Frobenius norm, so that

$$\mathbb{E}\left[\sup_{t\in[t_k,t_{k+1}]}\left\|\int_{t_k}^t K_s \mathrm{d}B_s\right\|^2\right] \leq hd \max_{t\in[t_k,t_{k+1}]} \mathbb{E}\left[\left\|A + \Sigma_\varepsilon^2 \nabla^2 \log p_{T-t}(\overleftarrow{\mathbf{U}}_t)\right\|^2\right]\,.$$

Using the L_t–Lipschitz continuous property of the score at time t we have

$$\mathbb{E}\left[\sup_{t\in[t_k,t_{k+1}]}\left\|\int_{t_k}^t K_s \mathrm{d}B_s\right\|^2\right] \leq 2hd\Big(\|A\|^2 + \max\{\sigma^4,\varepsilon^4\} \sup_{t\in[t_k,t_{k+1}]} L^2_{T-t_k}\Big). \tag{53}$$

Plugging (52) and (53) into (51) we obtain

$$A^2_{2,k} \leq C_a(\varepsilon) h^3 \tag{54}$$

with $C_a(\varepsilon)$ defined in (48).

Combining the bound on $A_{1,k}$ to the bound on $A_{2,k}$ yields,

$$\left\|\overleftarrow{\mathbf{U}}_{t_{k+1}} - \bar{\mathbf{U}}_{t_{k+1}}\right\|_{L_2} \leq \sqrt{1-h\delta_\varepsilon}\left\|\overleftarrow{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}\right\|_{L_2} + h\sqrt{h}C_a(\varepsilon)\,.$$

Using that $\overleftarrow{\mathbf{U}}_0 - \bar{\mathbf{U}}_0 = 0$ we have by induction

$$\begin{aligned}\left\|\overleftarrow{\mathbf{U}}_{t_N} - \bar{\mathbf{U}}_{t_N}\right\|_{L_2} &\leq \sum_{k=0}^{N-1}\prod_{j=k+1}^{N-1}\left(1-h\delta_\varepsilon\right)^{1/2} h\sqrt{h}C_a(\varepsilon)\,,\\ &\leq \frac{2}{\delta_\varepsilon}\sqrt{h}C_a(\varepsilon)\,,\end{aligned}$$

since $\sqrt{1-\delta_\varepsilon h} \leq 1 - h\delta_\varepsilon/2$. Letting $\Delta \to 0$ together with Fatou's lemma finishes the proof.

□

Lemma C.4 (Approximation error)**.** *Assume that H2′ and H3 hold. Then, there exists $\delta_\varepsilon > 0$ such that*

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{\mathbf{U}}^\infty_T\right), \mathcal{L}\left(\bar{\mathbf{U}}^\theta_T\right)\right) \leq \frac{2}{\delta_\varepsilon}\max\left\{\varepsilon^2, \sigma^2\right\} M\,.$$

Proof. Note that

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{\mathbf{U}}^\infty_T\right), \mathcal{L}\left(\bar{\mathbf{U}}^\theta_T\right)\right) \leq \left\|\bar{\mathbf{U}}^\infty_T - \bar{\mathbf{U}}^\theta_T\right\|_{L_2}\,.$$

Using a decomposition similar to that in C.3, with $t_N = T - \Delta$, we obtain:

$$\begin{aligned}&\left\|\bar{\mathbf{U}}^\infty_{t_{k+1}} - \bar{\mathbf{U}}^\theta_{t_{k+1}}\right\|_{L_2}\\ &= \left\|\bar{\mathbf{U}}^\infty_{t_k} - \bar{\mathbf{U}}^\theta_{t_k} + \int_{t_k}^{t_{k+1}} -A\left(\bar{\mathbf{U}}^\infty_{t_k} - \bar{\mathbf{U}}^\theta_{t_k}\right) + \Sigma^2_\varepsilon\left(\mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}^\infty_{t_k}\right) - s_\theta\left(T-t_k, \bar{\mathbf{U}}^\theta_{t_k}\right)\right)\mathrm{d}t\right\|_{L_2}\\ &\leq \left\|\bar{\mathbf{U}}^\infty_{t_k} - \bar{\mathbf{U}}^\theta_{t_k} + \int_{t_k}^{t_{k+1}} -A\left(\bar{\mathbf{U}}^\infty_{t_k} - \bar{\mathbf{U}}^\theta_{t_k}\right) + \Sigma^2_\varepsilon\left(\mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}^\infty_{t_k}\right) - \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}^\theta_{t_k}\right)\right)\mathrm{d}t\right\|_{L_2}\\ &\quad + \left\|\int_{t_k}^{t_{k+1}} \Sigma^2_\varepsilon\left(\mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}^\theta_{t_k}\right) - s_\theta\left(T-t_k, \bar{\mathbf{U}}^\theta_{t_k}\right)\right)\mathrm{d}t\right\|_{L_2}\\ &\leq B_{1,k} + B_{2,k}\,.\end{aligned}$$

For the first term, note that,

$$
\begin{aligned}
B_{1,k}^2 &= \left\|(\mathbf{I}_{2d} - hA)\left(\bar{\mathbf{U}}_{t_k}^{\infty} - \bar{\mathbf{U}}_{t_k}^{\theta}\right) + h\Sigma_{\varepsilon}^2\left(\mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}^{\infty}\right) - \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}^{\theta}\right)\right)\right\|_{L_2}^2 \\
&= \left\|(\mathbf{I}_{2d} - hA)\left(\bar{\mathbf{U}}_{t_k}^{\infty} - \bar{\mathbf{U}}_{t_k}^{\theta}\right)\right\|_{L_2}^2 + \left\|h\Sigma_{\varepsilon}^2\left(\mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}^{\infty}\right) - \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}^{\theta}\right)\right)\right\|_{L_2}^2 \\
&\qquad + 2h\mathbb{E}\left[\left(\bar{\mathbf{U}}_{t_k}^{\infty} - \bar{\mathbf{U}}_{t_k}^{\theta}\right)^{\top}(\mathbf{I}_{2d} - hA)^{\top}\Sigma_{\varepsilon}^2\left(\mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}^{\infty}\right) - \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}^{\theta}\right)\right)\right] .
\end{aligned}
$$

It follows that $B_{1,k}$ can be treated similarly to $A_{1,k}$. Using H2, H2′, and for h satisfying (50), we have

$$
B_{1,k} \leq \sqrt{1 - h\delta_{\varepsilon}}\left\|\bar{\mathbf{U}}_{t_k}^{\infty} - \bar{\mathbf{U}}_{t_k}^{\theta}\right\|_{L_2} ,
$$

where δ_{ε} is defined as in the proof of Lemma C.3. For $B_{2,k}$, using Assumption H3, we get

$$
B_{2,k} \leq h\left\|\Sigma_{\epsilon}\right\|^2 M \leq h\max\left\{\varepsilon^2, \sigma^2\right\} M .
$$

Finally, for h satisfying (50), it follows from the same argument as in the proof of Lemma C.3 that

$$
\left\|\bar{\mathbf{U}}_{t_N}^{\infty} - \bar{\mathbf{U}}_{t_N}^{\theta}\right\|_{L_2} \leq \frac{2}{\delta_{\varepsilon}}\max\left\{\varepsilon^2, \sigma^2\right\} M .
$$

Taking the limit as $\Delta \to 0$, toghether with Fatou's lemma finishes the proof. □

Lemma C.5 (Mixing time)**.** *Assume that H2′ holds. Then*

$$
\mathcal{W}_2\left(\mathcal{L}\left(\bar{\mathbf{U}}_T\right), \mathcal{L}\left(\bar{\mathbf{U}}_T^{\infty}\right)\right) \leq K_T e^{-aT}\mathcal{W}_2\left(\pi_{\mathrm{data}} \otimes \pi_v, \pi_{\infty}\right) ,
$$

with

$$
K_T := (1 + \max\{a+1; a(a+1)\}T) .
$$

Proof. Consider a synchronous coupling of the continuous-time interpolations $(\bar{\mathbf{U}}_t)_{t\in[0,T]}$ and $(\bar{\mathbf{U}}_t^{\infty})_{t\in[0,T]}$, with initialization

$$
\mathcal{W}_2\left(\pi_{\infty}, \mathcal{L}\left(\overrightarrow{\mathbf{U}}_T\right)\right) = \left\|\bar{\mathbf{U}}_0 - \bar{\mathbf{U}}_0^{\infty}\right\|_{L_2} .
$$

By definition of the $\mathcal{W}_2$ distance,

$$
\mathcal{W}_2\left(\mathcal{L}\left(\bar{\mathbf{U}}_T\right), \mathcal{L}\left(\bar{\mathbf{U}}_T^{\infty}\right)\right) \leq \left\|\bar{\mathbf{U}}_T - \bar{\mathbf{U}}_T^{\infty}\right\|_{L_2} .
$$

For $t \in [t_k, t_{k+1}]$ and with $t_N = T - \Delta$ we have that,

$$
\begin{aligned}
&\left\|\bar{\mathbf{U}}_{t_{k+1}} - \bar{\mathbf{U}}_{t_{k+1}}^{\infty}\right\|_{L_2} \\
&= \left\|\bar{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}^{\infty} + \int_{t_k}^{t_{k+1}} -A\left(\bar{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}^{\infty}\right) + \Sigma^2\left(\mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}\right) - \mathsf{s}_{T-t_k}\left(\bar{\mathbf{U}}_{t_k}^{\infty}\right)\right)\mathrm{d}t\right\|_{L_2} \\
&\leq \left\|\bar{\mathbf{U}}_{t_k} - \bar{\mathbf{U}}_{t_k}^{\infty}\right\|_{L_2}\delta_k ,
\end{aligned}
$$

where δ_k is defined as in (50). As a consequence,

$$
\left\|\bar{\mathbf{U}}_T - \bar{\mathbf{U}}_T^{\infty}\right\|_{L_2} \leq \left\|\bar{\mathbf{U}}_0 - \bar{\mathbf{U}}_0^{\infty}\right\|_{L_2}\prod_{\ell=0}^{N-1}\delta_{\ell} ,
$$

where we let $\Delta \to 0$ together with Fatou's lemma. Finally, using Lemma A.4, yields

$$
\left\|\bar{\mathbf{U}}_0 - \bar{\mathbf{U}}_0^{\infty}\right\|_{L_2} = \mathcal{W}_2\left(\pi_{\infty}, \mathcal{L}\left(\overrightarrow{\mathbf{U}}_T\right)\right) \leq K_T e^{-aT}\mathcal{W}_2\left(\pi_{\mathrm{data}} \otimes \pi_v, \pi_{\infty}\right) ,
$$

which finishes the proof. □

D Technical Lemmata

Lemma D.1. *Assume that H2 holds. Then, the data distribution $p_{\mathrm{data}}(x) \propto \exp(-(V(x)+H(x)))$ has sub-Gaussian tails,* i.e.*, there exist constants $C, \kappa > 0$ such that*

$$p_{\mathrm{data}}(x) \leq C\, \exp(-\kappa \|x\|^2)\,, \qquad x \in \mathbb{R}^d.$$

In particular, π_{data} admits finite moments of all orders.

Proof. By α–strong convexity of V, for all $x, y \in \mathbb{R}^d$,

$$V(x) \geq V(y) + \nabla V(y)^\top (x-y) + \frac{\alpha}{2}\|x-y\|^2\,.$$

Let x^* denote the unique minimizer of V, so that $\nabla V(x^*) = 0$. Then,

$$V(x) \geq V(x^*) + \frac{\alpha}{2}\|x-x^*\|^2 \geq \frac{\alpha}{4}\|x\|^2 - c_1\,,$$

for some constant $c_1 \in \mathbb{R}$. Since H is L–Lipschitz, we have

$$H(x) \geq H(x^*) - L\|x-x^*\| \geq -L\|x\| + c_2\,,$$

for some $c_2 \in \mathbb{R}$. Combining these two inequalities yields, for some $C \in \mathbb{R}$,

$$V(x) + H(x) \geq \frac{\alpha}{4}\|x\|^2 - L\|x\| + C\,.$$

Using Young's inequality $L\|x\| \leq \alpha\|x\|^2/8 + 2L^2/\alpha$, we obtain

$$V(x) + H(x) \geq \frac{\alpha}{8}\|x\|^2 - \frac{2L^2}{\alpha} + C\,.$$

Hence, up to a multiplicative constant,

$$p_{\mathrm{data}}(x) \;\propto\; \exp(-(V(x)+H(x))) \leq C'\, \exp\big(-\frac{\alpha}{8}\|x\|^2\big)\,,$$

for some $C' > 0$ which concludes the proof. □

Lemma D.2. *Assume that H2 holds and that there exist $m \in \mathbb{N}$ and $C > 0$ such that, for all $x \in \mathbb{R}^d$,*

$$\|\nabla V(x)\| \leq C\,(1 + \|x\|^m)\,. \tag{55}$$

Then, the relative Fisher Information between $\pi_0 = \pi_{\mathrm{data}} \otimes \pi_v$ (i.e. the initialization of the stochastic process defined in (4)*) and π_∞ is finite,* i.e.

$$\mathcal{I}(\pi_0|\pi_\infty) := \int \left\| \nabla \log\left(\frac{\mathrm{d}\pi_0}{\mathrm{d}\pi_\infty}(u)\right)\right\|^2 \pi_0(\mathrm{d}u) < \infty\,.$$

Proof. From Assumption H2, together with the fact that $\pi_0 = \pi_{\mathrm{data}} \otimes \pi_v$ and $\pi_v \sim \mathcal{N}\left(0_d, v^2\mathbf{I}_d\right)$,

$$p_0(u) = p_{\mathrm{data}}(x)\mathcal{N}(y; 0_d, v^2\mathbf{I}_d) \propto \mathrm{e}^{-(V(x)+H(x))}\mathrm{e}^{-\frac{\|y\|^2}{2v^2}}\,.$$

Therefore, the relative Fisher Information satisfies

$$\begin{aligned}
\mathcal{I}(\pi_0|\pi_\infty) &= \mathbb{E}\left[\left\| -\begin{pmatrix} \nabla V(\overrightarrow{X}_0) + \nabla H(\overrightarrow{X}_0) \\ v^{-2}\overrightarrow{V}_0 \end{pmatrix} + \Sigma_\infty^{-1}\begin{pmatrix} \overrightarrow{X}_0 \\ \overrightarrow{V}_0 \end{pmatrix}\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\| -\begin{pmatrix} \nabla V(\overrightarrow{X}_0) + \nabla H(\overrightarrow{X}_0) \\ v^{-2}\overrightarrow{V}_0 \end{pmatrix}\right\|^2\right] + 2\mathbb{E}\left[\left\|\Sigma_\infty^{-1}\begin{pmatrix} \overrightarrow{X}_0 \\ \overrightarrow{V}_0 \end{pmatrix}\right\|^2\right]\,.
\end{aligned}$$

By Lemma D.1, π_{data} has sub-Gaussian tails, hence

$$\mathbb{E}\left[\left\|\Sigma_\infty^{-1}\begin{pmatrix} \overrightarrow{X}_0 \\ \overrightarrow{V}_0 \end{pmatrix}\right\|^2\right] < \infty\,.$$

Moreover,

$$\mathbb{E}\left[\left\|-\begin{pmatrix}\nabla V(\overrightarrow{X}_0)+\nabla H(\overrightarrow{X}_0)\\ v^{-2}\overrightarrow{V}_0\end{pmatrix}\right\|^2\right] \le 2\mathbb{E}\left[\left\|\nabla V(\overrightarrow{X}_0)\right\|^2\right]+2\mathbb{E}\left[\left\|\nabla H(\overrightarrow{X}_0)\right\|^2\right]+v^{-4}\mathbb{E}\left[\left\|\overrightarrow{V}_0\right\|^2\right]$$

Since $\overrightarrow{V}_0$ is Gaussian, $\mathbb{E}[\|\overrightarrow{V}_0\|^2] < \infty$, and by Assumption H2, H is L-Lipschitz, so that $\mathbb{E}[\|\nabla H(\overrightarrow{X}_0)\|^2] \le L^2$. Using (55), there exist $m \in \mathbb{N}$ and $C > 0$ such that

$$\mathbb{E}\left[\left\|\nabla V(\overrightarrow{X}_0)\right\|\right] \le C\left(1+\mathbb{E}\left[\left\|\overrightarrow{X}_0\right\|^m\right]\right) < \infty\,,$$

using sub-Gaussianity of $\pi_{\rm data}$, which concludes the proof. □

Lemma D.3. *Assume that $(\overrightarrow{\mathbf{U}}_t)_{t\in[0,T]}$ is solution to* (9) *and that $\overrightarrow{\mathbf{U}}_0$ admits a second order moment, then for all $0 \le t \le T$, then for all $\varepsilon \ge 0$,*

$$\left\|\overleftarrow{\mathbf{U}}_t\right\|_{L_2}^2 \le \left(1+(a+1)^2(T-t)\right)^2 \mathrm{e}^{-2a(T-t)}\left\|\overrightarrow{\mathbf{U}}_0\right\|_{L_2}^2 + \frac{d}{2}\big(\sigma^2\max\{a,1/a\}+\frac{5\varepsilon^2}{a}\big) =: B\,. \tag{56}$$

Proof. Note that, in distribution,

$$\overrightarrow{\mathbf{U}}_t \overset{\mathcal{L}}{=} \mathrm{e}^{tA}\overrightarrow{\mathbf{U}}_0 + \Sigma_{0,t}^{1/2}G\,,$$

with $\overrightarrow{\mathbf{U}}_0 \sim \pi_{\rm data}\otimes\pi_v$, $G\sim\mathcal{N}(0,\mathbf{I}_{2d})$, and where G and $\overrightarrow{\mathbf{U}}_0$ are independent. Since G and $\overrightarrow{\mathbf{U}}_0$ are independent, using time-reversal and sub-multiplicativity of matrix norms, we have that

$$\begin{aligned}\mathbb{E}\left[\left\|\overleftarrow{\mathbf{U}}_{T-t}\right\|^2\right] = \mathbb{E}\left[\left\|\overrightarrow{\mathbf{U}}_t\right\|^2\right] &= \mathbb{E}\left[\left\|\mathrm{e}^{tA}\overrightarrow{\mathbf{U}}_0\right\|^2\right]+\mathbb{E}\left[\left\|\Sigma_{0,t}^{1/2}G\right\|^2\right]\\ &\le \left\|\mathrm{e}^{tA}\right\|^2\mathbb{E}\left[\left\|\overrightarrow{\mathbf{U}}_0\right\|^2\right]+\left\|\Sigma_{0,t}^{1/2}\right\|^2\mathbb{E}\left[\|G\|^2\right]\\ &= \left\|\mathrm{e}^{tA}\right\|^2\mathbb{E}\left[\left\|\overrightarrow{\mathbf{U}}_0\right\|^2\right]+2d\lambda_{\max}(\Sigma_{0,t})\,.\end{aligned}$$

We conclude by applying Lemma A.1 to bound $\left\|\mathrm{e}^{tA}\right\|^2$ and Lemma A.3 to bound $\lambda_{\max}(\Sigma_{0,t})$. □

Remark D.4. Lemma D.3 holds true when $\overleftarrow{\mathbf{U}}_t$ is defined as in (4) by setting $\varepsilon = 0$.

Lemma D.5. *Assume that $(\overrightarrow{\mathbf{U}}_t)_{t\in[0,T]}$ is solution to* (9)*, then,*

$$\mathbb{E}\left[\left\|\mathsf{s}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)\right\|^2\right] \le \frac{2d}{\lambda_{\min}(\Sigma_{0,T-t})}\,,$$

where $\Sigma_{0,t}$ is defined in (20)*.*

Proof. By the time-reversal property,

$$\mathbb{E}\left[\left\|\mathsf{s}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)\right\|^2\right] = \mathbb{E}\left[\left\|\mathsf{s}_{T-t}\left(\overrightarrow{\mathbf{U}}_{T-t}\right)\right\|^2\right]\,.$$

Note that

$$\mathsf{s}_{T-t}(\overrightarrow{\mathbf{U}}_{T-t}) = \mathbb{E}\left[\Sigma_{0,T-t}^{-1}(e^{(T-t)A}\overrightarrow{\mathbf{U}}_0-\overrightarrow{\mathbf{U}}_{T-t})|\overrightarrow{\mathbf{U}}_{T-t}\right]\,,$$

then, using Jensen's inequality and the tower property,

$$\mathbb{E}\left[\left\|\mathsf{s}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)\right\|^2\right] \le \mathbb{E}\left[\left\|\Sigma_{0,T-t}^{-1}\left(e^{(T-t)A}\overrightarrow{\mathbf{U}}_0-\overrightarrow{\mathbf{U}}_{T-t}\right)\right\|^2\right]\,.$$

Since $\overrightarrow{\mathbf{U}}_t \overset{\mathcal{L}}{=} \mathrm{e}^{tA}\overrightarrow{\mathbf{U}}_0+\Sigma_{0,t}^{1/2}G$ with $\overrightarrow{\mathbf{U}}_0\sim\pi_{\rm data}\otimes\pi_v$, $G\sim\mathcal{N}(0,\mathbf{I}_{2d})$, and where G and $\overrightarrow{\mathbf{U}}_0$ are independent, we have

$$\mathbb{E}\left[\left\|\mathsf{s}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)\right\|^2\right] \le \mathbb{E}\left[\left\|\Sigma_{0,T-t}^{-1/2}G\right\|^2\right] = \mathrm{Tr}\left(\Sigma_{0,T-t}^{-1}\right)\,,$$

which completes the proof.

□

Lemma D.6. *Assume that* $(\overleftarrow{\mathbf{U}}_t)_{t\in[0,T]}$ *is solution to the backward SDE associated with* (9)*. Then,*

$$\mathrm{d}(\nabla \log p_{T-t}(\overleftarrow{\mathbf{U}}_t)) = A^\top \nabla \log p_{T-t}(\overleftarrow{\mathbf{U}}_t)\mathrm{d}t + \nabla^2 \log p_{T-t}(\overleftarrow{\mathbf{U}}_t)\Sigma_\varepsilon \mathrm{d}B_t\,.$$

Proof. The Fokker-Plank equation for the SDE defined in (4) yields, for $u \in \mathbb{R}^{2d}$,

$$\partial_t p_t(u) = -\mathrm{div}(Aup_t(u)) + \frac{1}{2}\mathrm{div}(\Sigma_\varepsilon^2 \nabla p_t(u))\,. \tag{57}$$

First, using the notation introduced in (10),

$$\begin{aligned}
\mathrm{div}(Aup_t(u)) &= \sum_{i=1}^{2d} \frac{\partial Aup_t(u)}{\partial u_i} \\
&= \sum_{i=1}^{2d}\sum_{j=1}^{2d} \frac{\partial}{\partial u_i} A_{ij}u_j p_t(u) \\
&= \sum_{i=1}^{2d} A_{ii}p_t(u) + \sum_{i=1}^{2d}\sum_{j=1}^{2d} A_{ij}u_j \frac{\partial}{\partial_i} p_t(u) \\
&= \sum_{i=1}^{2d} A_{ii}p_t(u) + (Au)^\top \nabla p_t(u) \\
&= \mathrm{Tr}(A)p_t(u) + (Au)^\top \nabla p_t(u) \\
&= p_t(u)\left(\mathrm{Tr}(A) + (Au)^\top \mathsf{s}_t(u)\right)\,.
\end{aligned}$$

Second, using the product rule for divergence,

$$\begin{aligned}
\frac{1}{2}\mathrm{div}(\Sigma_\varepsilon^2 \nabla p_t(u)) &= \frac{1}{2}\mathrm{div}(\Sigma_\varepsilon^2 p_t(u)\mathsf{s}_t(u)) \\
&= \frac{1}{2}\mathrm{div}(p_t(u)\Sigma_\varepsilon^2 \mathsf{s}_t(u)) \\
&= \frac{1}{2}\left(p_t(u)\mathrm{div}(\Sigma_\varepsilon^2 \mathsf{s}_t(u)) + (\Sigma_\varepsilon^2 \mathsf{s}_t(u))^\top \nabla p_t(u)\right) \\
&= \frac{1}{2}p_t(u)\left(\mathrm{div}(\Sigma_\varepsilon^2 \mathsf{s}_t(u)) + \mathsf{s}_t(u)^\top \Sigma_\varepsilon^2 \mathsf{s}_t(u)\right)\,.
\end{aligned}$$

Hence, dividing (57) by p_t yields

$$\partial_t \log p_t(u) = -\mathrm{Tr}(A) - (Au)^\top \mathsf{s}_t(u) + \frac{1}{2}\left[\left(\mathrm{div}(\Sigma_\varepsilon^2 \mathsf{s}_t(u)) + \mathsf{s}_t(u)^\top \Sigma_\varepsilon^2 \mathsf{s}_t(u)\right)\right]\,,$$

so that,

$$\partial_t \log p_{T-t}(u) = \mathrm{Tr}(A) + (Au)^\top \mathsf{s}_{T-t}(u) - \frac{1}{2}\left[\left(\mathrm{div}(\Sigma_\varepsilon^2 \mathsf{s}_{T-t}(u)) + \mathsf{s}_{T-t}(u)^\top \Sigma_\varepsilon^2 \mathsf{s}_{T-t}(u)\right)\right]\,.$$

Recall that the backward process can be written as

$$\mathrm{d}\overleftarrow{\mathbf{U}}_t = (-A\overleftarrow{\mathbf{U}}_t + \Sigma_\varepsilon^2 \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t))\mathrm{d}t + \Sigma_\varepsilon \mathrm{d}B_t\,.$$

Hence, by Itô's formula,

$$\begin{aligned}
\mathrm{d}(\mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)) &= \partial_t \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\mathrm{d}t + \nabla \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\mathrm{d}U_t + \frac{1}{2}\mathrm{Tr}\left(\Sigma_\varepsilon^2 \nabla^2 \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\right)\mathrm{d}t \\
&= \nabla \partial_t \log p_{T-t}(\overleftarrow{\mathbf{U}}_t)\mathrm{d}t - \nabla \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)A\overleftarrow{\mathbf{U}}_t \mathrm{d}t + \nabla \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\Sigma_\varepsilon^2 \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\mathrm{d}t \\
&\quad + \frac{1}{2}\mathrm{Tr}\left(\Sigma_\varepsilon^2 \nabla^2 \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\right)\mathrm{d}t + \nabla \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\Sigma_\varepsilon \mathrm{d}B_t \\
&= \nabla\left(\partial_t \log p_{T-t}(\overleftarrow{\mathbf{U}}_t) + \frac{1}{2}\mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)^\top \Sigma_\varepsilon^2 \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t) + \frac{1}{2}\mathrm{div}(\Sigma_\varepsilon^2 \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t))\right)dt \\
&\quad - \nabla \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)A\overleftarrow{\mathbf{U}}_t \mathrm{d}t + \nabla \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\Sigma_\varepsilon \mathrm{d}B_t \\
&= A^\top \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\mathrm{d}t + \nabla \mathsf{s}_{T-t}(\overleftarrow{\mathbf{U}}_t)\Sigma_\varepsilon \mathrm{d}B_t\,,
\end{aligned}$$

which completes the proof and where we used that for $u \in \mathbb{R}^{2d}$, $2\nabla^2 \log p_t(u)\Sigma_\varepsilon^2 \mathsf{s}_t(u) = \nabla(\mathsf{s}_t(u)^\top \Sigma_\varepsilon^2 \mathsf{s}_t(u))$ and $\nabla \mathrm{div}(\Sigma_\varepsilon^2 \mathsf{s}_t(u)) = \nabla \mathrm{Tr}(\Sigma_\varepsilon^2 \nabla^2 \mathsf{s}_t(u))$. Indeed, for $k \in \{1, ..., 2d\}$, with $g(u) = \nabla \log p_t(u)$, and therefore $g_i(u) = \frac{\partial}{\partial u_i} g(u)$

$$\begin{aligned}
\frac{\partial}{\partial u_k}\left(\nabla g(u)^\top \Sigma_\varepsilon^2 \nabla g(u)\right) &= \frac{\partial}{\partial u_k} \sum_{i,j} g_i(u) \Sigma_{\varepsilon,ij}^2 g_j(u) \\
&= \sum_{i,j} \Sigma_{\varepsilon,ij}^2 \left(g_j(u) \frac{\partial}{\partial u_k} g_i(u) + g_i(u) \frac{\partial}{\partial u_k} g_j(u) \right) \\
&= 2 \sum_{i=1}^{2d} \Sigma_{\varepsilon,ii}^2 \left(g_i(u) \frac{\partial}{\partial u_k} g_i(u) \right) \\
&= 2 \sum_{i=1}^{2d} \Sigma_{\varepsilon,ii}^2 \left(\frac{\partial}{\partial u_i} g(u) \frac{\partial}{\partial u_k} \frac{\partial}{\partial u_i} g(u) \right) \\
&= \left[2\nabla^2 g(u) \Sigma_\varepsilon^2 \nabla g(u) \right]_k .
\end{aligned}$$

□

Lemma D.7. *Assume that* $(\overleftarrow{\mathbf{U}}_t)_{t\in[0,T]}$ *is solution to the backward SDE associated with* (9)*. Then,*

$$\mathrm{d}(\tilde{\mathsf{s}}_{T-t}(\overleftarrow{\mathbf{U}}_t)) = -\tilde{A}_\epsilon^\top \tilde{\mathsf{s}}_{T-t}(\overleftarrow{\mathbf{U}}_t)\mathrm{d}t + \nabla^2 \log \tilde{p}_{T-t}(\overleftarrow{\mathbf{U}}_t)\Sigma_\epsilon \mathrm{d}B_t \,. \tag{58}$$

Proof. Recall that p_∞ is the stationary distribution of (4) so that using Fokker-Planck equation we get, for $u \in \mathbb{R}^{2d}$,

$$\begin{aligned}
0 = &-\mathrm{Tr}(A) - (Au)^\top \nabla \log p_\infty(u) \\
&+ \frac{1}{2}\left[\mathrm{div}\left(\Sigma^2 \nabla \log p_\infty(u)\right) + \nabla \log p_\infty(u)^\top \Sigma^2 \nabla \log p_\infty(u) \right] .
\end{aligned}$$

Using that $\tilde{p}_t = p_t / p_\infty$, and Fokker-Planck as in Lemma D.6

$$\begin{aligned}
\partial_t \log \tilde{p}_t(u) = &-(Au)^\top \tilde{\mathsf{s}}_t(u) \\
&+ \frac{1}{2}\left[\mathrm{div}\left(\Sigma^2 \tilde{\mathsf{s}}_t(u)\right) + \tilde{\mathsf{s}}_t(u)^\top \Sigma^2 \tilde{\mathsf{s}}_t(u) \right] \\
&+ \tilde{\mathsf{s}}_t(u)^\top \Sigma^2 \nabla \log p_\infty(u) \,.
\end{aligned}$$

Using the definition of $\tilde{A}_\epsilon$, we have,

$$\partial_t \log \tilde{p}_t(u) = (\tilde{A}_\epsilon u)^\top \tilde{\mathsf{s}}_t(u) + \frac{1}{2}\left[\mathrm{div}\left(\Sigma^2 \tilde{\mathsf{s}}_t(u)\right) + \tilde{\mathsf{s}}_t(u)^\top \Sigma^2 \tilde{\mathsf{s}}_t(u) \right] ,$$

and therefore,

$$\partial_t \log \tilde{p}_{T-t}(u) = -(\tilde{A}_\epsilon u)^\top \tilde{\mathsf{s}}_{T-t}(u) - \frac{1}{2}\left[\mathrm{div}\left(\Sigma^2 \tilde{\mathsf{s}}_{T-t}(u)\right) + \tilde{\mathsf{s}}_{T-t}(u)^\top \Sigma^2 \tilde{\mathsf{s}}_{T-t}(u) \right] .$$

Recall that the modified backward process can be written as

$$\mathrm{d}\overleftarrow{\mathbf{U}}_t = (\tilde{A}_\epsilon \overleftarrow{\mathbf{U}}_t + \Sigma^2 \tilde{\mathsf{s}}_{T-t}(\overleftarrow{\mathbf{U}}_t))\mathrm{d}t + \Sigma_\epsilon dB_t \,.$$

Hence, by Itô's formula,

$$\begin{aligned}
&\mathrm{d}(\tilde{\mathsf{s}}_{T-t}(\overleftarrow{\mathbf{U}}_t)) \\
&= \partial_t \tilde{\mathsf{s}}_{T-t}(\overleftarrow{\mathbf{U}}_t)\mathrm{d}t + \nabla^2 \log \tilde{p}_{T-t}(\overleftarrow{\mathbf{U}}_t) d\overleftarrow{\mathbf{U}}_t + \frac{1}{2}\mathrm{Tr}\left(\Sigma^2 \nabla^2 \tilde{\mathsf{s}}_{T-t}(\overleftarrow{\mathbf{U}}_t)\right)\mathrm{d}t \\
&= \nabla\Big(\partial_t \log \tilde{p}_{T-t}(\overleftarrow{\mathbf{U}}_t)\mathrm{d}t + \frac{1}{2}\tilde{\mathsf{s}}_{T-t}(\overleftarrow{\mathbf{U}}_t)^\top \Sigma^2 \tilde{\mathsf{s}}_{T-t}(\overleftarrow{\mathbf{U}}_t) + \frac{1}{2}\mathrm{div}\left(\Sigma^2 \tilde{\mathsf{s}}_{T-t}(\overleftarrow{\mathbf{U}}_t)\right)\Big) \\
&+ \nabla^2 \log \tilde{p}_{T-t}(\overleftarrow{\mathbf{U}}_t)\tilde{A}_\epsilon \overleftarrow{\mathbf{U}}_t \mathrm{d}t + \nabla^2 \log \tilde{p}_{T-t}(\overleftarrow{\mathbf{U}}_t)\Sigma_\epsilon \mathrm{d}B_t \\
&= -\tilde{A}_\epsilon^\top \tilde{\mathsf{s}}_{T-t}(\overleftarrow{\mathbf{U}}_t) + \nabla^2 \log \tilde{p}_{T-t}(\overleftarrow{\mathbf{U}}_t)\Sigma_\epsilon \mathrm{d}B_t \,,
\end{aligned}$$

which completes the proof.

□

Lemma D.8. *Let Δ be an arbitrary fixed positive constant, and assume that $(\overleftarrow{\mathbf{U}}_t)_{t\in[0,T-\Delta]}$ is the solution to* (11)*. Then, there exists a universal constant $C > 0$ such that*

$$\mathbb{E}\left[\left\|\nabla \log \tilde{p}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right) - \nabla \log \tilde{p}_{T-t_k}\left(\overleftarrow{\mathbf{U}}_{t_k}\right)\right\|^2\right] \leq C\left(g(t_{k+1}) - g(t_k)\right) ,$$

for $t \in [t_k, t_{k+1}]$, with

$$g(t) := \mathbb{E}\left[\left\|\nabla \log \tilde{p}_{T-t}\left(\overleftarrow{\mathbf{U}}_t\right)\right\|^2\right] . \tag{59}$$

Proof. The argument follows from an adaptation of Conforti et al. (Proposition 3.2, 2025) to our setting. Let $Y_t := \nabla \log \tilde{p}_{T-t}(\overleftarrow{\mathbf{U}}_t)$ and $Z_t := \nabla^2 \log \tilde{p}_{T-t}(\overleftarrow{\mathbf{U}}_t)$. From (58), the process $(Y_t)_{t\in[0,T]}$ satisfies

$$\mathrm{d}Y_t = -\tilde{A}_\epsilon^\top Y_t \mathrm{d}t + Z_t \Sigma_\epsilon \mathrm{d}B_t .$$

Applying Itô's formula to $\|Y_t\|^2$ yields

$$\mathrm{d}\left\|Y_t\right\|^2 = -2\langle Y_t, \tilde{A}_\epsilon^\top Y_t\rangle \,\mathrm{d}t + 2\langle Y_t, Z_t\Sigma_\epsilon \,\mathrm{d}B_t\rangle + \|Z_t\Sigma_\epsilon\|_{\mathrm{Fr}}^2 \,\mathrm{d}t .$$

Therefore, there exists a constant $c > 0$, depending only on a, such that

$$\mathrm{d}\left\|Y_t\right\|^2 \geq c\left(\left\|Y_t\right\|^2 + \|Z_t\Sigma_\epsilon\|_{\mathrm{Fr}}^2\right)\mathrm{d}t + \tilde{H}_t \mathrm{d}B_t ,$$

where $\tilde{H}_t$ denotes a stochastic process. Moreover, following the argument of Conforti et al. (Lemma 3.3, 2025), the stochastic integral $\int_0^t \tilde{H}_r \mathrm{d}B_r$ is a true martingale. Using this and integrating over $[t_k, t]$, we deduce that there exists a universal constant $C > 0$ (whose value may change in the course of the argument) such that

$$\mathbb{E}\left[\left\|Y_t - Y_{t_k}\right\|^2\right] \leq C\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\|Y_s\right\|^2 + \|Z_s\Sigma_\epsilon\|_{\mathrm{Fr}}^2\right]\mathrm{d}s \leq C(g(t_{k+1}) - g(t_k)) .$$

□

Lemma D.9. *Let $A \in \mathbb{R}^{n\times n}$ be an invertible matrix, and let $B \in \mathbb{R}^{n\times n}$ be such that $A - B$ is also invertible. Then,*

$$(A-B)^{-1} - A^{-1} = (A-B)^{-1}BA^{-1}.$$

Proof. Note that

$$(A-B)^{-1} - A^{-1} = (A-B)^{-1}AA^{-1} - A^{-1} = \left[(A-B)^{-1}A - \mathbf{I}_n\right]A^{-1} ,$$

and

$$(A-B)^{-1}A = (A-B)^{-1}\left((A-B)+B\right) = \mathbf{I}_n + (A-B)^{-1}B ,$$

so that

$$\left[(A-B)^{-1}A - \mathbf{I}_n\right]A^{-1} = (A-B)^{-1}BA^{-1} ,$$

which completes the proof. □

E Numerical Illustration

This section provides additional details on the numerical implementation described in Section 4.

E.1 CLD training and sampling

Algorithms 1 and 2 show the training and sampling procedures for the CLD-based approaches, respectively.

Algorithm 1 CLD Training

Require: Dataset $\mathcal{D}$, batch size B, network $s_\theta(\cdot, t)$, a positive weight function $\lambda : [0, T] \to \mathbb{R}_+$ and $\epsilon \geq 0$.

Precompute $\tilde{\Sigma}_{0,t} = \Sigma_{0,t} + e^{tA}\mathrm{diag}(0\mathbf{I}_d, v^2\mathbf{I}_d)(e^{tA})^\top$. ▷ The value of $\Sigma_{0,t}$ depends on ϵ, see Lemma A.2. (eq 23).
while not converged **do**
  Sample $\{x^{(i)}\}_{i=1}^B \sim \mathcal{D}$
  Sample $\{t^{(i)}\}_{i=1}^B \sim \mathcal{U}([0, T])$
  Sample $\{\varepsilon^{(i)}\}_{i=1}^B \sim \mathcal{N}(0, \mathbf{I}_{2d})$
  $\overrightarrow{\mathbf{U}}_{t^{(i)}} = e^{t^{(i)}A}\left(\overrightarrow{X}_0, 0_d\right)^\top + (\tilde{\Sigma}_{0,t^{(i)}})^{1/2}\varepsilon^{(i)}$
  $\mathcal{L} \leftarrow \frac{1}{B}\sum_{i=1}^B \lambda(t^{(i)}) \left\| s_\theta\left(t^{(i)}, \overrightarrow{\mathbf{U}}_{t^{(i)}}\right) + (\tilde{\Sigma}_{0,t^{(i)}})^{-1/2}\varepsilon^{(i)} \right\|^2$
  Update θ by taking gradient step on $\nabla_\theta \mathcal{L}$
end while

Algorithm 2 CLD Sampling

Require: Learned network s_θ, number of discretization steps N and $\epsilon \geq 0$.

$h \leftarrow T/N$
$\bar{\mathbf{U}}_0 \sim \pi_\infty$
for $k = 0$ down to $N - 1$ **do**
  $t_k \leftarrow k\,h$
  Sample $Z_k \sim \mathcal{N}(0, \mathbf{I}_{2d})$ ▷ π_∞ depends on ϵ, see (21) in Lemma A.2.
  $\bar{\mathbf{U}}^\theta_{t_{k+1}} = \bar{\mathbf{U}}^\theta_{t_k} + h\left(\tilde{A}_\epsilon \bar{\mathbf{U}}^\theta_{t_k} + \Sigma_\epsilon^2 s_\theta(t_k, \bar{\mathbf{U}}^\theta_{t_k})\right) + \sqrt{h}\Sigma_\epsilon Z_k$
end for
return First d coordinates of $\bar{\mathbf{U}}^\theta_{t_N}$ ▷ Return position only, discard velocity.

E.2 Time-rescaling of the forward SDE

Following Dockhorn et al. (2022), one often implements in practice a time-rescaled version of

$$\mathrm{d}\overrightarrow{\mathbf{U}}_t = A\overrightarrow{\mathbf{U}}_t\mathrm{d}t + \Sigma_\epsilon \mathrm{d}B_t\,,$$

by introducing a positive noise schedule $\beta\colon [0, 1] \to [0, \infty)$ and setting

$$\overrightarrow{\tilde{\mathbf{U}}}_t = \overrightarrow{\mathbf{U}}_{\tau(t)} \quad \text{and} \quad \tau(t) = \int_0^t \beta(s)\mathrm{d}s\,.$$

Equivalently, $\overrightarrow{\tilde{\mathbf{U}}}_t$ satisfies the inhomogeneous SDE

$$\mathrm{d}\overrightarrow{\tilde{\mathbf{U}}}_t = \underbrace{\beta(t)A}_{=\tilde{A}(t)}\overrightarrow{\tilde{\mathbf{U}}}_t\mathrm{d}t + \underbrace{\sqrt{\beta(t)}\,\Sigma_\epsilon}_{=\tilde{\Sigma}_\epsilon(t)}\mathrm{d}B_t\,,$$

In the critically-damped example (Equation (4)), we have

$$\tilde{A}(t) = \begin{pmatrix} 0 & \beta(t)\,a^2 \\ -\beta(t) & -2a\,\beta(t) \end{pmatrix} \otimes \mathbf{I}_d, \quad \tilde{\Sigma}_\epsilon(t) = \sqrt{\beta(t)}\Sigma_\epsilon \otimes \mathbf{I}_d\,.$$

Mean factor. Since $\overrightarrow{\tilde{\mathbf{U}}}_t = \overrightarrow{\mathbf{U}}_{\tau(t)}$, we can deduce from the homogeneous solution the mean factor,

$$\mathbb{E}\left[\overrightarrow{\tilde{\mathbf{U}}}_t \mid \overrightarrow{\mathbf{U}}_0\right] = e^{-a\,\tau(t)} \left(\begin{pmatrix} 1 + a\,\tau(t) & a^2\,\tau(t) \\ -\tau(t) & 1 - a\,\tau(t) \end{pmatrix} \otimes \mathbf{I}_d \right) \overrightarrow{\mathbf{U}}_0 \,.$$

Covariance. Again by the time-change $\tau(t)$, one has

$$\mathrm{Cov}\big(\overrightarrow{\tilde{\mathbf{U}}}_t \mid \overrightarrow{\mathbf{U}}_0\big) = \mathrm{Cov}\big(\overrightarrow{\mathbf{U}}_{\tau(t)} \mid \overrightarrow{\mathbf{U}}_0\big) = \int_0^{\tau(t)} e^{sA} \Sigma_\epsilon \, \Sigma_\epsilon^T e^{sA^T} \mathrm{d}s \,.$$

Affine schedule. A popular and simple choice of noise schedule is an affine noise schedule given by

$$\beta(t) = \beta_1 t + \beta_0 \,, \quad \tau(t) = \frac{\beta_1}{2} t^2 + \beta_0 t \,.$$

E.3 Score approximation

Denoising Score Matching (DSM). Recall that the conditional score function of the forward process (4) given the initial data distribution is Gaussian,

$$\nabla \log p_t(\overrightarrow{\mathbf{U}}_t | \overrightarrow{\mathbf{U}}_0) = -\Sigma_{0,t}^{-1} \left(\overrightarrow{\mathbf{U}}_t - e^{tA} \overrightarrow{\mathbf{U}}_0 \right) .$$

Hence, following Vincent (2011) the conditional denoising score matching loss $\mathcal{L}_{\mathrm{cond}}$, for $\theta \in \Theta$, $s_\theta(t,x) : [0,T] \times \mathbb{R}^{2d} \mapsto \mathbb{R}^{2d}$ and $Z_{2d} \sim \mathcal{N}(0, \mathbf{I}_{2d})$ can be written as

$$\begin{aligned} \mathcal{L}_{\mathrm{DSM}}(\theta) &= \mathbb{E}\left[\lambda(t) \left\| s_\theta\left(\tau, \overrightarrow{\mathbf{U}}_\tau\right) - \nabla \log p_\tau \left(\overrightarrow{\mathbf{U}}_\tau | \overrightarrow{\mathbf{U}}_0 \right) \right\|^2 \right] \\ &= \mathbb{E}\left[\lambda(t) \left\| s_\theta\left(\tau, \mathrm{e}^{\tau A} \overrightarrow{\mathbf{U}}_0 + \sqrt{\Sigma_{0,\tau}} Z_{2d}\right) + \Sigma_{0,t}^{-1/2} Z_{2d} \right\|^2 \right] , \end{aligned}$$

where $\tau \sim \mathcal{U}\,[0,T]$, $\tau \perp Z_{2d}$ and $\lambda : [0,T] \mapsto \mathbb{R}_{>0}$.

Hybrid Score Matching (HSM). It has been shown in Dockhorn et al. (2022) that another loss, potentially more stable numerically can be obtained by conditioning only on $\overrightarrow{X}_0$ rather than on the full state $\overrightarrow{\mathbf{U}}_0 = (\overrightarrow{X}_0, \overrightarrow{V}_0)^\top$. This hybrid score matching loss can be derived by marginalizing out the velocity component $\overrightarrow{V}_0 \sim \mathcal{N}\left(0_d, v^2 \mathbf{I}_d\right)$, $\overrightarrow{V}_0 \perp \overrightarrow{X}_0$ in the conditional score function,

$$\begin{aligned} \mathcal{L}_{\mathrm{HSM}}(\theta) &= \mathbb{E}\left[\lambda(t) \left\| s_\theta(\tau, \overrightarrow{\mathbf{U}}_\tau) - \nabla \log p_\tau(\overrightarrow{\mathbf{U}}_\tau \mid \overrightarrow{X}_0) \right\|^2 \right] \\ &= \mathbb{E}\left[\lambda(t) \left\| s_\theta\left(\tau, e^{\tau A} \begin{pmatrix} \overrightarrow{X}_0 \\ 0_d \end{pmatrix} + \sqrt{\Sigma'_{0,\tau}} Z_{2d}\right) + (\Sigma'_{0,\tau})^{-1/2} Z_{2d} \right\|^2 \right] , \end{aligned}$$

with $Z_{2d} \sim \mathcal{N}(0, \mathbf{I}_{2d})$ independent of $\tau \sim \mathcal{U}[0,T]$ and

$$\Sigma'_{0,\tau} = \Sigma_{0,\tau} + e^{\tau A} \begin{pmatrix} 0 & 0 \\ 0 & v^2 \mathbf{I}_d \end{pmatrix} (e^{\tau A})^\top .$$

E.4 Neural network architectures

In Figure 3 we detail the neural network used in the illustration. The input layer is composed of a vector x in dimension $2d$ and the time t. Both are respectively embedded using a linear transformation or a sine/cosine transformation (Nichol and Dhariwal, 2021) of width `mid_features`. Then, 3 dense layers of constant width `mid_features` followed by SiLu activations and skip connections regarding the time embedding. The output layer is linear resulting in a vector of dimension d (when $\varepsilon = 0$) and $2d$ (when $\varepsilon \neq 0$).

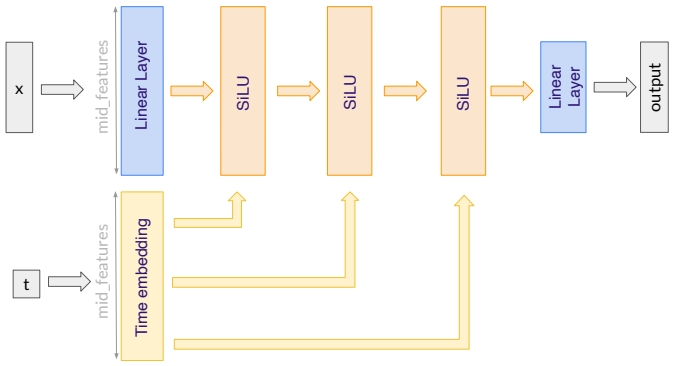

Figure 3: Neural network architecture.

E.5 Additional experiments

We present additional experimental results for the MG25 distribution in dimension 100 and the 2D-diamond dataset. The MG25 distribution is defined as a Gaussian mixture model with 25 modes in dimension 100, defined as

$$\pi_{\text{data}}(x) = \frac{1}{25} \sum_{(j,k)\in\{-2,\dots,2\}^2} \varphi_{\mu_{jk},\Sigma_d}(x)$$

with $\varphi_{\mu_{jk},\Sigma_d}$ denoting the probability density function of the Gaussian distribution with covariance matrix $\Sigma_d = \text{diag}\,(0.01, 0.01, 0.1, ..., 0.1)$ and mean vector $\mu_{jk} = [j, k, 0, 0, 0..., 0]^\top$. This dataset has been previously used in Thin et al. (2021); Strasman et al. (2025). The 2D-diamond distribution is a two-dimensional dataset with well-separated modes, used as a synthetic dataset in Dockhorn et al. (2022).

Tables 1, 2 and 3 report the sliced-Wasserstein error for different values of the regularization parameter $\varepsilon \in \{0, 0.1, 0.25, 0.5, 1\}$ and drift coefficient $a \in \{0.1, 0.25, 0.5, 1, 2\}$, using the same experimental setup as for the Funnel dataset described in Section 4. Both tables 1 and 2 highlight the improvement in generation quality achieved with smaller regularization values of ε. Table 3 report the values displayed in Figure 1 with the associated standard deviations.

Table 1: Comparison of mean Wasserstein distance for different noise levels ε on the MG25-100D (mean $\pm$ standard deviation across 5 runs; lower is better).

ε	$a = 0.1$	$a = 0.25$	$a = 0.5$	$a = 1.0$	$a = 2.0$
0	0.284 ± 0.002	0.199 ± 0.001	0.034 ± 0.002	0.009 ± 0.001	0.009 ± 0.001
0.1	0.192 ± 0.001	0.159 ± 0.001	0.026 ± 0.001	$\mathbf{0.005 \pm 0.001}$	0.008 ± 0.001
0.25	$\mathbf{0.013 \pm 0.001}$	0.065 ± 0.001	0.015 ± 0.001	0.007 ± 0.001	$\mathbf{0.007 \pm 0.001}$
0.5	0.191 ± 0.007	$\mathbf{0.004 \pm 0.001}$	$\mathbf{0.009 \pm 0.001}$	0.008 ± 0.001	0.008 ± 0.001
1	0.389 ± 0.030	0.045 ± 0.003	0.011 ± 0.002	0.006 ± 0.001	0.008 ± 0.001

Table 2: Comparison of mean Wasserstein distance for different noise levels ε on the Diamond-2D (mean $\pm$ standard deviation across 5 runs; lower is better).

ε	$a = 0.1$	$a = 0.25$	$a = 0.5$	$a = 1.0$	$a = 2.0$
0	0.322 ± 0.001	0.256 ± 0.004	0.039 ± 0.002	0.007 ± 0.001	0.007 ± 0.002
0.1	0.234 ± 0.001	0.198 ± 0.003	0.026 ± 0.004	0.004 ± 0.001	0.005 ± 0.001
0.25	$\mathbf{0.048 \pm 0.001}$	0.074 ± 0.003	0.021 ± 0.002	$\mathbf{0.004 \pm 0.001}$	$\mathbf{0.005 \pm 0.001}$
0.5	0.073 ± 0.002	$\mathbf{0.008 \pm 0.001}$	$\mathbf{0.008 \pm 0.002}$	0.006 ± 0.002	0.006 ± 0.002
1	0.095 ± 0.002	0.029 ± 0.002	0.014 ± 0.001	0.013 ± 0.001	0.011 ± 0.001

Table 3: Comparison of mean Wasserstein distance for different noise levels ε on the Funnel-100D (mean $\pm$ standard deviation across 5 runs; lower is better).

ε	$a = 0.1$	$a = 0.25$	$a = 0.5$	$a = 1.0$	$a = 2.0$
0	0.991 ± 0.001	0.73 ± 0.002	0.291 ± 0.005	0.225 ± 0.056	0.223 ± 0.011
0.1	0.705 ± 0.001	0.632 ± 0.002	0.278 ± 0.001	0.158 ± 0.027	0.198 ± 0.004
0.25	$\mathbf{0.277 \pm 0.002}$	0.409 ± 0.003	0.248 ± 0.012	$\mathbf{0.137 \pm 0.005}$	$\mathbf{0.179 \pm 0.006}$
0.5	1.171 ± 0.015	$\mathbf{0.248 \pm 0.002}$	0.228 ± 0.005	0.157 ± 0.002	0.203 ± 0.003
1	2.885 ± 0.016	0.785 ± 0.011	$\mathbf{0.191 \pm 0.008}$	0.253 ± 0.006	0.233 ± 0.002

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract summarizes the key contributions of the paper, clarifying their position relative to existing work and highlighting improvements over the state of the art.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The authors are transparent about the nature of the assumptions and mathematical models they study, and consistently highlight these choices as part of the scope and limitations of their work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Each theoretical result contains references to assumptions clearly defined in the main document. The rigorous proofs are also provided in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All the numerical experiments are well documented either in the main body text or on the appendices. The contributions are primarily theoretical and methodological, and are supported by numerical experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper will be accompanied of a public github repository with scripts and notebooks to reproduce the numerical experiments. For the submission, the codes are provided as a supplementary zip file.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We refer to the appendices for exhaustive implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All the numerical experiments are run several times, so that the performance metrics are given as averages over the different runs, with associated standard deviations. This information is available for instance graphically with confidence regions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Although the paper is theoretical, it provides information about the computational resources used in all experiments, including the types of compute workers (CPU or GPU).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research does not involve any human subjects or participants and contains no (sensitive) data.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This paper is purely methodological/theoretical. Results are established for pre-existing algorithms and their variants. Thus, there is no specific positive/negative societal impacts of this research.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper is theoretical and does not involve the release of data or models that have a high risk for misuse. Therefore, no safeguards were necessary.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: No existing assets were used, so no credits or licenses apply.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code is provided and well-documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not applicable

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We did not employ any LLMs during any stage of the research process.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.