Dimension-Uniform Discretization Analysis of Preconditioned Annealed Langevin Dynamics for Multimodal Gaussian Mixtures

Lorenzo Baldassari
University of Basel
lorenzo.baldassari@unibas.ch

Josselin Garnier
Ecole Polytechnique, IP Paris
josselin.garnier@polytechnique.edu

Knut Sølna
University of California Irvine
ksolna@uci.edu

Maarten V. de Hoop
Rice University
mvd2@rice.edu

Abstract

Obtaining stable diffusion-based samplers in high- and infinite-dimensional settings is challenging because errors can accumulate across high-frequency coordinates and make the dynamics unstable under refinement of the finite-dimensional approximation of the underlying function-space problem. Discretization is a typical source of such errors, and preconditioning with a suitable spectral decay is one way to control their accumulation. In this paper, we study this problem for preconditioned annealed Langevin dynamics (ALD) applied to Gaussian mixtures. We first show that Euler–Maruyama (EM) discretization, by treating the stiff linear part of the annealed score with a forward Euler step, imposes a stability constraint coupling the preconditioner with the annealed covariance scale. Together with the conditions ensuring dimension-uniform control of the annealed dynamics, this constraint forces the initial smoothed law to remain uniformly close to the target across dimensions. We then consider an exponential-integrator scheme that integrates the stiff linear part of the annealed score exactly. Under explicit spectral summability conditions coupling the smoothing covariance, the component covariance spectra, and the preconditioner, we prove a dimension-uniform Kullback–Leibler (KL) bound for this scheme. This bound can be made arbitrarily small, uniformly in dimension, by allowing enough time for annealing and then refining the time mesh accordingly. Importantly, these conditions allow regimes in which the KL divergence between the target and the initial smoothed law diverges with dimension, showing that the restrictions imposed by EM are scheme-dependent rather than intrinsic to ALD.

1 Introduction

Many sampling problems are infinite-dimensional: the target is a probability measure on a function space, while computation requires a finite-dimensional approximation [48, 14, 25]. Designing diffusion-based samplers whose behavior remains stable as this approximation is refined is therefore crucial. The difficulty, and a central challenge in infinite-dimensional sampling, is that errors that are harmless at a fixed truncation level can accumulate across high-frequency coordinates. Consequently, a sampler may appear stable at any fixed dimension while failing to provide guarantees that are uniform in dimension. Preconditioning is one way to address this challenge. A preconditioner modifies the geometry of the diffusion; in a diagonal basis, this amounts to rescaling the dynamics across spectral directions. Chosen appropriately, it can improve the behavior of the sampler as the finite-dimensional approximation is refined [42, 14]. However, it introduces a trade-off: a less damped

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

preconditioner can accelerate motion in directions where the dynamics would otherwise be slow, but once errors are present, the same preconditioner may amplify their high-frequency components. Thus the choice of preconditioner depends not only on the target and on the continuous-time dynamics, but also on the error sources introduced by the algorithm. Discretization is one such source.

This paper studies this issue for preconditioned annealed Langevin dynamics (ALD), a sampling procedure that applies Langevin dynamics along a path of intermediate distributions [23]. These distributions are "annealed": they interpolate between an initially smoothed version of the target and the target itself, by gradually removing the smoothing [28, 21, 38]. The goal is to mitigate the metastability and slow mixing that standard Langevin dynamics can exhibit on multimodal targets, especially in high dimensions [44, 34, 18]: rather than targeting a complex distribution directly, ALD first explores a smoother landscape and is then progressively guided toward the original one.

Despite its empirical success [46, 47, 51], the theoretical understanding of ALD is still developing [10]. In finite dimensions, [23] has shown that annealing can provide provable advantages over standard Langevin dynamics for multimodal sampling, including polynomial-in-dimension mixing guarantees. In infinite dimensions, however, far fewer guarantees are available. One key difficulty is that ALD does not exactly follow the prescribed annealing path: its law may deviate from the intermediate annealed distributions, and this annealing-induced bias must be controlled uniformly under refinement of the finite-dimensional approximations. For the continuous-time dynamics, [2] has shown that preconditioned ALD targeting Gaussian mixtures—a classical and widely used family for approximating multimodal distributions [36, 22]—can be controlled uniformly in dimension under spectral conditions involving the smoothing covariance, the component covariances, and the preconditioner, and that this control is robust to perturbations of the annealed score. It does not, however, address a question central to implementation:

Under which conditions on the smoothing, component-covariance, and preconditioning spectra does this uniform-in-dimension behavior persist after time discretization?

In this paper, we answer this question in the same multimodal setting. To isolate the effect of time discretization, we assume exact access to the annealed scores and exact initialization from the initial smoothed law, leaving aside the perturbative regime studied in the continuous-time analysis of [2].

Our first step is to understand which restrictions are introduced by the most standard discretization, Euler–Maruyama (EM) [29]. In Section 3, we show that treating the linear part of the annealed score by a forward Euler step imposes a stability condition coupling the preconditioner with the annealed covariance scale. In high-frequency coordinates, where this scale is small, the linear drift can be large (the "stiff" directions), so stability requires strong high-frequency damping of the preconditioner. Combined with the assumptions used to keep the annealing bias uniformly controlled, this stability condition further forces the initial smoothed law to remain uniformly close to the target across dimensions. We first prove this for an explicit bimodal family, where the relevant quantities can be computed directly (Proposition 3.2); we then show that the same implication holds in a more general setting under the conservative spectral conditions used to control the annealing bias (Proposition 3.3).

Since the difficulty with EM in infinite dimensions comes from the way it discretizes the stiff linear part of the annealed score, one may expect that alternative schemes treating this part differently could lead to more favorable conditions. In Section 4, we study such an alternative: an exponential-integrator scheme [45, 30, 50] of the type suggested by [23] in their finite-dimensional analysis of ALD. We call it an exact-linear-part (ELP) scheme because, at each time step, it integrates the stiff linear part of the annealed score exactly and freezes the remaining nonlinear term. Our main result (Theorem 4.1) proves that ELP admits a dimension-uniform KL bound under explicit spectral summability conditions on the smoothing, component covariances, and the preconditioner, which can be less restrictive than those for EM: in particular, they do not force the initial smoothed law to remain uniformly close to the target across dimensions; see Proposition 4.1 and the numerical illustrations of Section 5. These conditions also clarify how the preconditioner should be chosen: it must be strong enough to keep the annealing bias under control, but sufficiently damped in the tail to prevent discretization errors from accumulating across high-frequency coordinates. Appendix C makes this trade-off explicit in a power-law regime, yielding admissible preconditioning spectra.

Related Work. Annealing-based strategies have a long history in sampling [21, 35, 37, 38]. The basic idea is to replace the direct sampling problem by a sequence of easier intermediate problems,

starting from a distribution that is easy to sample from and gradually transforming it into the target. If consecutive distributions are sufficiently close, approximate samples from one stage can provide a warm start for the next, thereby easing the transition toward the target. When this principle is implemented through a Langevin diffusion driven by a time-dependent score schedule, one obtains ALD [46, 47, 51, 52].

Recent work has begun to develop theoretical guarantees for ALD, both from the sampling and generative-modeling perspectives [23, 11, 10, 13]. These analyses, however, do not address whether the behavior of ALD remains stable under successive finite-dimensional approximations of the infinite-dimensional target. Available discretization analyses are likewise not uniform in dimension: for instance, [13] analyzes an Euler–Maruyama discretization of ALD with dimension-dependent bounds, whereas [23] is closer in spirit to our approach and uses an exponential-integrator discretization [45, 30, 50], but still obtains bounds with polynomial dependence on dimension.

Recently, [2] addressed the infinite-dimensional problem by studying a preconditioned version of ALD. This analysis is closest to ours in spirit: it shows how dimension-uniform control can be obtained from the spectral properties of the annealing geometry and the preconditioner, and identifies conditions under which score and initialization perturbations remain compatible with such robustness. That analysis, however, is limited to continuous time. Our paper shows that the dimension-uniform control proved there is not automatically inherited by time discretization: as the finite-dimensional approximation of the underlying infinite-dimensional target is refined, Euler–Maruyama imposes a restrictive high-frequency condition that the exact-linear-part exponential integrator avoids.

Finally, our analysis connects naturally to the classical function-space Markov chain Monte Carlo literature, which develops sampling schemes designed to yield dimension-independent behavior [14, 25, 15, 5]. It also contributes to a growing line of work emphasizing the role of preconditioning in improving the efficiency and stability of sampling dynamics [9, 41, 6, 26, 43, 16, 31, 32]. Given the close connection between ALD and score-based diffusion models, our work further relates to recent efforts to analyze these models in function space [27, 19, 3, 40, 8, 4, 33, 24, 20, 1], although here we take a purely sampling perspective and assume access to the score.

2 Continuous-Time Background

In this section, we briefly recall the infinite-dimensional framework of [2], which provides the modeling assumptions and truncation regime used throughout the paper. We then introduce the continuous-time preconditioned annealed Langevin dynamics at the core of our analysis.

Infinite-Dimensional Setting and Truncation Regime. Let H be a separable Hilbert space with orthonormal basis $(e_j)_{j\geq 1}$, and consider the infinite-dimensional Gaussian mixture

$$\rho_\star^\infty = \sum_{i\in I} w_i \mathcal{N}(m_i, \Sigma_i),$$

where I is finite or countable, $w_i > 0$, and $\sum_{i\in I} w_i = 1$. We work in a diagonal setting,

$$m_i = \sum_{j\geq 1} m_{ij} e_j, \qquad \Sigma_i e_j = \sigma_{ij} e_j, \qquad \sigma_{ij} > 0,$$

and assume the finite-energy condition

$$\sum_{j\geq 1} m_{ij}^2 < \infty, \qquad \sum_{j\geq 1} \sigma_{ij} < \infty, \qquad i \in I,$$

so that $m_i \in H$ and Σ_i is trace class. Hence each component $\mathcal{N}(m_i, \Sigma_i)$ is well defined on H [7].

In this paper, rather than sampling $\rho_\star^\infty$ directly, we study its successive finite-dimensional truncations. For each $d \geq 1$, let P_d denote the orthogonal projection onto $\mathrm{span}\{e_1, \ldots, e_d\}$, and define

$$\rho_\star^d := (P_d)_\# \rho_\star^\infty = \sum_{i\in I} w_i \mathcal{N}(m_i^d, \Sigma_i^d),$$

where $m_i^d := (m_{i1}, \ldots, m_{id})$, $\Sigma_i^d := \mathrm{Diag}(\sigma_{i1}, \ldots, \sigma_{id})$. The diagonal formulation makes the refinement structure explicit: increasing d adds new coordinates in a fixed basis while leaving the

previously retained coordinates unchanged. This is the regime in which we study dimension-uniform control. Importantly, this assumption is less restrictive than it may first appear: at each fixed truncation level, diagonal Gaussian mixtures remain highly expressive and, with sufficiently many components, can approximate general mixtures in KL distance.

For the discretization analysis developed in Sections 3 and 4, we will also use the quantities

$$\underline{\sigma}_j := \inf_{i\in I} \sigma_{ij}, \qquad \overline{\sigma}_j := \sup_{i\in I} \sigma_{ij}, \qquad \overline{m}_j := \sup_{i\in I} |m_{ij}|, \qquad j \geq 1,$$

and we assume that $0 < \underline{\sigma}_j \leq \overline{\sigma}_j < \infty$ and $\overline{m}_j < \infty$ for all $j \geq 1$.

Annealed Langevin Dynamics. We now recall the continuous-time annealed Langevin dynamics (ALD) studied in [2]. For a time horizon $T > 0$, we define for $t \in [0, T]$:

$$\kappa_t := \frac{T-t}{T}, \qquad \rho_t^d = \rho_\star^d * \mathcal{N}(0, \kappa_t C^d), \quad \text{with } C^d = \mathrm{Diag}(\lambda_1, \ldots, \lambda_d),$$

and we assume $\sum_{j\geq 1} \lambda_j < \infty$, so that the smoothing covariance is trace class and the corresponding Gaussian perturbation is H-valued. Then ρ_t^d interpolates between the smoothed law $\rho_0^d = \rho_\star^d * \mathcal{N}(0, C^d)$ and the final target $\rho_T^d = \rho_\star^d$.

Starting from $X_0^d \sim \rho_0^d$, consider the preconditioned annealed Langevin diffusion

$$dX_t^d = \Gamma^d \nabla \log \rho_t^d(X_t^d)\, dt + \sqrt{2\Gamma^d}\, dW_t^d, \qquad t \in [0, T], \tag{1}$$

where W_t^d is a standard Brownian motion in $\mathbb{R}^d$ and

$$\Gamma^d = \mathrm{Diag}(\gamma_1, \ldots, \gamma_d)$$

is a diagonal preconditioner, whose role will be discussed later. We denote the terminal law by

$$\rho_T^{\mathrm{ALD},d} := \mathrm{Law}(X_T^d).$$

Note that (1) does not target $\rho_\star^d$ directly: its drift uses the time-dependent score $\nabla \log \rho_t^d$ rather than the final score $\nabla \log \rho_\star^d$. Accordingly, its law does not in general coincide with the annealing path, and one may have $\rho_T^{\mathrm{ALD},d} \neq \rho_\star^d$. We quantify this *annealing-induced bias* via the KL divergence

$$\mathcal{B}_{\mathrm{ann}}^d(T) := \mathrm{KL}(\rho_\star^d \,\|\, \rho_T^{\mathrm{ALD},d}).$$

A natural question is whether one can choose a single dimension-uniform T so that $\mathcal{B}_{\mathrm{ann}}^d(T)$ remains small over successive refinements $(\rho_\star^d)_{d\geq 1}$ of the infinite-dimensional target. An explicit answer was recently given in [2], through the following theorem.

Theorem 2.1. *Let*

$$\mathcal{K}_d := \frac{1}{16} \sum_{i\in I} w_i \sum_{j=1}^{d} \frac{\lambda_j}{\gamma_j} \log\Big(1 + \frac{\lambda_j}{\sigma_{ij}}\Big).$$

If $\sup_{d\geq 1} \mathcal{K}_d < \infty$*, then for every* $\varepsilon > 0$*, choosing*

$$T_\varepsilon = \varepsilon^{-1} \sup_{d\geq 1} \mathcal{K}_d,$$

one has

$$\mathrm{KL}(\rho_\star^d \,\|\, \rho_{T_\varepsilon}^{\mathrm{ALD},d}) \leq \varepsilon, \qquad \textit{for all } d \geq 1.$$

A convenient sufficient condition for $\sup_d \mathcal{K}_d < \infty$ is obtained from $\log(1+u) \leq u$:

$$\sum_{i\in I} w_i \sum_{j\geq 1} \frac{\lambda_j^2}{\gamma_j \sigma_{ij}} < \infty. \tag{2}$$

Hence, dimension-uniform control reduces to a spectral compatibility condition between the smoothing (λ_j), the preconditioner (γ_j), and the target covariance spectra (σ_{ij}).

On the Role of the Preconditioner in Continuous Time. At first sight, Theorem 2.1 may suggest that preconditioning is a *cost-free accelerator*: the annealing-bias bound depends on Γ^d only through the ratios λ_j/γ_j, so it seems that making the preconditioner larger can only help. One might even be tempted to read the theorem as saying that by choosing rapidly growing preconditioner coefficients (γ_j) one can make the required continuous-time horizon arbitrarily small, essentially without trade-off. This interpretation is misleading. Theorem 2.1 is idealized: it assumes exact access to the annealed score and exact initialization from ρ_0^d, and therefore isolates only the annealing-bias contribution. The robustness analysis in [2] shows that, once score perturbations or initialization mismatch are introduced, sufficient decay of (γ_j) helps prevent high-frequency errors from accumulating across coordinates and destroying dimension-uniform control. This is the main lesson we import from the continuous-time analysis. Since time discretization also introduces coordinatewise errors, one should expect a similar phenomenon there as well.

The remainder of the paper makes this point explicit. We begin with the most classical discretization scheme, Euler–Maruyama [29]. We show that, as the finite-dimensional approximations of the infinite-dimensional target are refined, Euler–Maruyama requires strong high-frequency damping of the preconditioner. This motivates the alternative scheme analyzed in Section 4.

3 High-Frequency Stiffness in Euler–Maruyama Discretization

Here, we consider the Euler–Maruyama (EM) discretization of the ALD diffusion. We focus on a specific limitation arising from the fact that EM applies a forward Euler step to the linear part of the annealed score. Along high-frequency coordinates, where the annealed covariance scale is small, this linear part can have large drift coefficients; stability then requires strong high-frequency damping of the preconditioner. Combined with the spectral condition used to control the annealing-induced bias in (2), this stability restriction forces a regime in which the smoothed law used to initialize ALD must remain uniformly close to the target across dimensions.

In this section, we continue to work in the diagonal Gaussian-mixture setting of Section 2. We write

$$D_{i,t}^d := \Sigma_i^d + \kappa_t C^d.$$

Since the annealed score is

$$\nabla \log \rho_t^d(x) = \sum_{i\in I} p_{i,t}^d(x)\Big(-(D_{i,t}^d)^{-1}(x-m_i^d)\Big), \text{ with responsibilities } p_{i,t}^d(x) = \frac{w_i \varphi_{i,t}^d(x)}{\sum_{k\in I} w_k \varphi_{k,t}^d(x)},$$

and $\varphi_{i,t}^d$ is the probability density function of the distribution $\mathcal{N}(m_i^d, D_{i,t}^d)$, it is useful to rewrite it as

$$\nabla \log \rho_t^d(x) = -(B_t^d)^{-1}x + G_t^d(x), \tag{3}$$

where $B_t^d := \mathrm{Diag}(b_{t,1}, \dots, b_{t,d})$, $b_{t,j} := \underline{\sigma}_j + \kappa_t \lambda_j$, and

$$G_t^d(x) := \sum_{i\in I} p_{i,t}^d(x)\Big(\big((B_t^d)^{-1} - (D_{i,t}^d)^{-1}\big)x + (D_{i,t}^d)^{-1} m_i^d\Big).$$

This decomposition isolates a diagonal linear part, $-(B_t^d)^{-1}x$, whose high-frequency coefficients can become large, from the nonlinear mixture correction $G_t^d(x)$.

With a constant step size $h > 0$ and grid points $t_n = nh$, the EM scheme reads

$$Y_{n+1}^d = Y_n^d - h\,\Gamma^d (B_{t_n}^d)^{-1} Y_n^d + h\,\Gamma^d G_{t_n}^d(Y_n^d) + \sqrt{2h\,\Gamma^d}\,\xi_n, \tag{4}$$

where $\xi_n \sim \mathcal{N}(0, I_d)$ are i.i.d. Here, the linear and nonlinear terms are treated in the same way: both are frozen at time t_n and advanced by a forward Euler step. For the linear part, this yields coordinate-wise the factor $1 - h\gamma_j/b_{t_n,j}$. Stability requires this factor to have modulus at most one, giving the following immediate condition.

Proposition 3.1. *The coordinate-wise linear factors in* (4) *satisfy* $\big|1 - h\gamma_j/b_{t_n,j}\big| \le 1$ *for all mesh indices and all* $j \ge 1$ *if and only if*

$$h \sup_{n,j} \frac{\gamma_j}{\underline{\sigma}_j + \kappa_{t_n}\lambda_j} \le 2.$$

Proposition 3.1 highlights two possible regimes. If $\lambda_j \gtrsim \underline{\sigma}_j$ along the tail, then $b_{t_0,j} = \underline{\sigma}_j + \lambda_j \asymp \lambda_j$, so the condition above yields $\sup_{j\geq 1} \gamma_j/\lambda_j < \infty$. Together with the annealing-bias condition (2), this forces

$$\sum_{i\in I} w_i \sum_{j\geq 1} \frac{\lambda_j}{\sigma_{ij}} < \infty.$$

This case is very restrictive. For example, it fails whenever there exists a component i_0 with $\sigma_{i_0 j} \asymp \underline{\sigma}_j$ along the tail, and in particular whenever the mixture is finite. Therefore, the regime that is more relevant for our analysis is the more favorable one

$$\lambda_j \lesssim \underline{\sigma}_j.$$

Under this condition, Proposition 3.1 implies that uniform control of the EM linear factors with a fixed step size requires

$$\sup_{j\geq 1} \frac{\gamma_j}{\underline{\sigma}_j} < \infty. \tag{5}$$

In the successive-refinement regimes of interest, where $\underline{\sigma}_j \to 0$ as $j \to \infty$, this EM stability condition rules out any non-decaying, let alone increasing, preconditioner. The next proposition shows what this restriction entails. Its proof is given in Appendix A.1. In a simple symmetric bimodal family, uniform-in-dimension control of the annealing-induced bias together with (5) forces the smoothed law used to initialize ALD to remain uniformly close to the target across dimensions.

Proposition 3.2. *Let e_1^d be the first canonical basis vector of $\mathbb{R}^d$. Consider the symmetric bimodal target $\pi_\star^d = \frac{1}{2}\mathcal{N}(ae_1^d, \Sigma^d) + \frac{1}{2}\mathcal{N}(-ae_1^d, \Sigma^d)$, where $a \neq 0$ is fixed and $\Sigma^d = \mathrm{Diag}(\sigma_1, \ldots, \sigma_d)$. Let $\pi_t^d = \pi_\star^d * \mathcal{N}(0, \kappa_t C^d)$, with $C^d = \mathrm{Diag}(\lambda_1, \ldots, \lambda_d)$ and $\Gamma^d = \mathrm{Diag}(\gamma_1, \ldots, \gamma_d)$. Let $(X_t^d)_{t\in[0,T]}$ be the corresponding ALD diffusion initialized from $\pi_0^d = \pi_\star^d * \mathcal{N}(0, C^d)$.*

Assume the EM stability condition $\sup_{j\geq 2} \gamma_j/\sigma_j < \infty$. If, for some fixed $T > 0$,

$$\sup_{d\geq 1} \mathrm{KL}(\pi_\star^d \,\|\, \mathrm{Law}(X_T^d)) < \infty,$$

then

$$\sup_{d\geq 1} \mathrm{KL}(\pi_\star^d \,\|\, \pi_0^d) < \infty.$$

If one now returns to the general diagonal Gaussian-mixture setting and combines the EM stability condition with the annealing-bias condition (2), the same conclusion holds. Indeed, assume the EM stability condition $\sup_{j\geq 1} \gamma_j/\underline{\sigma}_j < \infty$, then there exists $M < \infty$ such that, for every i and j, $\gamma_j/\sigma_{ij} \leq \gamma_j/\underline{\sigma}_j \leq M$, and hence $\lambda_j^2/\sigma_{ij}^2 \leq M\, \lambda_j^2/(\gamma_j \sigma_{ij})$. Summing in i and j and using (2) gives

$$\sum_{i\in I} w_i \sum_{j\geq 1} \frac{\lambda_j^2}{\sigma_{ij}^2} < \infty. \tag{6}$$

As the next proposition shows, with its proof in Appendix A.2, this forces the initial annealed law to remain uniformly close to the target across dimensions.

Proposition 3.3. *If (6) holds, then $\sup_{d\geq 1} \mathrm{KL}(\rho_\star^d \,\|\, \rho_0^d) < \infty$.*

The conclusions arising from Proposition 3.1 are, however, *scheme-dependent*. They do not rule out dimension-uniform time discretization of ALD. Rather, they show that the additional restrictions imposed by EM come from approximating the stiff diagonal linear part by a forward Euler step. This motivates the exponential-integrator discretization [45, 30, 50] studied in the next section. On each discretization interval $[t_n, t_{n+1})$, the scheme integrates the stiff linear part of the annealed score exactly while freezing the remaining nonlinear term at its value at the left endpoint t_n.

4 Dimension-Uniform Exact-Linear-Part Discretization

We now define this scheme and analyze its dimension-uniform behavior.

Definition of the Scheme. Recall from the previous section that the annealed score decomposes as

$$\nabla \log \rho_t^d(x) = -(B_t^d)^{-1}x + G_t^d(x),$$

with a time-dependent diagonal linear part $-(B_t^d)^{-1}x$ and a nonlinear mixture correction $G_t^d(x)$, where $B_t^d = \mathrm{Diag}(b_{t,1}, \dots, b_{t,d})$, $b_{t,j} = \underline{\sigma}_j + \kappa_t \lambda_j$. Now fix a mesh

$$0 = t_0 < t_1 < \cdots < t_N = T, \qquad h_n := t_{n+1} - t_n, \qquad h_{\max} := \max_{0 \le n \le N-1} h_n.$$

The key idea is to freeze only the nonlinear correction over each interval $[t_n, t_{n+1})$, while integrating the resulting time-dependent diagonal linear SDE exactly. We call this coordinate-wise exponential-integrator discretization the exact-linear-part (ELP) scheme.

Definition 4.1. *The ELP discretization of the ALD diffusion is the sequence $(Y_n^d)_{n=0}^N$ defined by $Y_0^d \sim \rho_0^d$ and*

$$Y_{n+1,j}^d = \phi_{n,j} Y_{n,j}^d + \psi_{n,j} G_{t_n,j}^d(Y_n^d) + \xi_{n,j}, \qquad j = 1, \dots, d, \tag{7}$$

where

$$\phi_{n,j} := \exp\left(-\int_{t_n}^{t_{n+1}} \frac{\gamma_j}{b_{s,j}}\, ds\right), \qquad \psi_{n,j} := \int_{t_n}^{t_{n+1}} \exp\left(-\int_s^{t_{n+1}} \frac{\gamma_j}{b_{r,j}}\, dr\right) \gamma_j\, ds,$$

and $\xi_n = (\xi_{n,1}, \dots, \xi_{n,d})$ has independent centered Gaussian coordinates with

$$\mathrm{Var}(\xi_{n,j}) = 2\gamma_j \int_{t_n}^{t_{n+1}} \exp\left(-2\int_s^{t_{n+1}} \frac{\gamma_j}{b_{r,j}}\, dr\right) ds.$$

In the symmetric bimodal family of Proposition 3.2, one has

$$\phi_{n,j} = \exp\left(-\gamma_j \int_{t_n}^{t_{n+1}} \frac{dt}{\sigma_j + \kappa_t \lambda_j}\right) = \left(\frac{\sigma_j + \kappa_{t_{n+1}}\lambda_j}{\sigma_j + \kappa_{t_n}\lambda_j}\right)^{\gamma_j T/\lambda_j} \in (0,1).$$

Thus the linear factor is contractive for every coefficient and over every time-step interval, without imposing the EM stability condition on $\gamma_j/\underline{\sigma}_j$. This already suggests that the restrictions imposed by EM in Section 3 are scheme-dependent rather than intrinsic to ALD.

A Dimension-Uniform KL Bound. We are now in the position to state the main dimension-uniform KL estimate for the ELP scheme. For this purpose, it is convenient to view the discrete scheme through its continuous-time interpolation. For $t \in [t_n, t_{n+1})$, let Y_t^d solve

$$dY_t^d = -\Gamma^d (B_t^d)^{-1} Y_t^d\, dt + \Gamma^d G_{t_n}^d(Y_{t_n}^d)\, dt + \sqrt{2\Gamma^d}\, dW_t^d, \qquad Y_{t_n}^d = Y_n^d. \tag{8}$$

Then $Y_T^d = Y_N^d$, so $\mathrm{Law}(Y_T^d)$ is the terminal law of the ELP scheme. The proof compares this interpolation with an auxiliary process $\widehat{X}^d$ whose marginals satisfy $\mathrm{Law}(\widehat{X}_t^d) = \rho_t^d$; see (17)–(18). By data processing and Girsanov's formula, the terminal KL is bounded by a $(\Gamma^d)^{-1/2}$-weighted path-space energy of the drift mismatch over $[0,T]$. Along the auxiliary path, this mismatch splits into the transport velocity of the annealing path, whose energy is $J_{\mathrm{ann}}^d(T)$ in (19), and the freezing defect produced by replacing G_t^d with $G_{t_n}^d$, defined in (20). This yields the KL estimate below; the assumptions are spelled out after the theorem and verified in Appendix B.1.

Theorem 4.1. *Assume that the conditions* (9)–(15) *below hold. Then there exists a dimension-independent constant C_{disc} (that depends only on the bounds in* (9)–(15)*) such that*

$$\sup_{d \ge 1} \mathrm{KL}\big(\rho_\star^d \,\|\, \mathrm{Law}(Y_T^d)\big) \le \frac{1}{8T} \sum_{j \ge 1} \frac{\lambda_j^2}{\gamma_j \underline{\sigma}_j} + C_{\mathrm{disc}}(1+T^2) h_{\max}.$$

Thus, for every $\varepsilon_{\mathrm{ann}}, \varepsilon_{\mathrm{disc}} > 0$, if

$$T \ge \frac{1}{8\varepsilon_{\mathrm{ann}}} \sum_{j \ge 1} \frac{\lambda_j^2}{\gamma_j \underline{\sigma}_j}, \qquad h_{\max} \le \frac{\varepsilon_{\mathrm{disc}}}{C_{\mathrm{disc}}(1+T^2)},$$

then we have

$$\sup_{d \ge 1} \mathrm{KL}\big(\rho_\star^d \,\|\, \mathrm{Law}(Y_T^d)\big) \le \varepsilon_{\mathrm{ann}} + \varepsilon_{\mathrm{disc}}.$$

This bound separates two effects. The first term is the annealing contribution: it decreases as the annealing time T increases. The second term is the discretization contribution: after T is fixed, it can be made small by refining the time mesh. Crucially, under the spectral assumptions (9)–(15), this trade-off is uniform in the truncation dimension d.

We now spell out the conditions (9)–(15) that constitute one set of sufficient conditions. They are organized according to the quantities controlled in the proof given in Appendix B.1.

- The first group controls moments of the annealing path, the linear drift $-\Gamma^d(B_t^d)^{-1}x$, and the weighted growth of the nonlinear correction G_t^d. The moment condition

$$\sum_{j\geq 1}(\overline{\sigma}_j + \lambda_j + \overline{m}_j^2) < \infty \tag{9}$$

  is used to obtain uniform moment bounds for the annealing path; see Lemma B.1. The linear-drift condition

$$\sum_{j\geq 1}\gamma_j < \infty, \qquad \sum_{j\geq 1}\gamma_j^2\frac{\overline{\sigma}_j + \lambda_j + \overline{m}_j^2}{\underline{\sigma}_j^2} < \infty \tag{10}$$

  controls the fourth moment of the diagonal linear drift; see Lemma B.3. Finally,

$$\sum_{j\geq 1}\gamma_j\frac{(\overline{\sigma}_j - \underline{\sigma}_j)^2 + \underline{\sigma}_j^2\overline{m}_j^2}{\underline{\sigma}_j^4} < \infty \tag{11}$$

  controls the weighted growth of the nonlinear correction G_t^d; see Lemma B.5. Together, these estimates also enter the fourth-moment bound for the full auxiliary drift; see Lemma B.6.

- The second group controls the freezing defect produced by replacing the current nonlinear field G_t^d by the frozen field $G_{t_n}^d$ on each time interval. Under the auxiliary path law, the defect is $G_t^d(\widehat{X}_t^d) - G_{t_n}^d(\widehat{X}_{t_n}^d)$, for $t \in [t_n, t_{n+1})$. We split it into a spatial increment and a temporal increment:

$$G_t^d(\widehat{X}_t^d) - G_{t_n}^d(\widehat{X}_{t_n}^d) = \big(G_t^d(\widehat{X}_t^d) - G_t^d(\widehat{X}_{t_n}^d)\big) + \big(G_t^d(\widehat{X}_{t_n}^d) - G_{t_n}^d(\widehat{X}_{t_n}^d)\big).$$

  Since $G_t^d(x) = \sum_{i\in I} p_{i,t}^d(x)c_{i,t}^d(x)$, the estimates require controlling the spatial and temporal variation of both the affine corrections $c_{i,t}^d$ and the responsibilities $p_{i,t}^d$.
  For the spatial increment, we assume

$$\sum_{j\geq 1}\frac{(\overline{\sigma}_j - \underline{\sigma}_j)^2}{\underline{\sigma}_j^4}(\overline{\sigma}_j + \lambda_j + \overline{m}_j^2) < \infty, \qquad \sum_{j\geq 1}\sup_{i,\ell\in I}\frac{|m_{ij} - m_{\ell j}|^2}{\underline{\sigma}_j^2} < \infty. \tag{12}$$

  These conditions control the spatial variation of the affine corrections and of the responsibilities; see Lemma B.8 and Lemma B.9.
  For the time variation of the affine corrections, we assume

$$\sum_{j\geq 1}\frac{\lambda_j(\overline{\sigma}_j - \underline{\sigma}_j)}{\underline{\sigma}_j^3}(\overline{\sigma}_j + \lambda_j + \overline{m}_j^2) < \infty, \qquad \sum_{j\geq 1}\gamma_j\frac{\lambda_j^2(\overline{\sigma}_j - \underline{\sigma}_j)^2}{\underline{\sigma}_j^6}(\overline{\sigma}_j + \lambda_j + \overline{m}_j^2) < \infty. \tag{13}$$

  These conditions enter the bounds on the time derivatives of the affine correction coefficients and the temporal freezing defect; see Lemma B.4 and Lemma B.11.
  For the time variation of the responsibilities, we assume

$$\sum_{j\geq 1}\frac{\lambda_j^2}{\underline{\sigma}_j^2}\left(\sup_{i,\ell\in I}\frac{|m_{ij} - m_{\ell j}|^2}{\underline{\sigma}_j^2} + \frac{\overline{m}_j^2(\overline{\sigma}_j - \underline{\sigma}_j)^2}{\underline{\sigma}_j^4}\right) < \infty, \qquad \sum_{j\geq 1}\gamma_j\lambda_j^2\frac{\overline{m}_j^2}{\underline{\sigma}_j^4} < \infty. \tag{14}$$

  These conditions control the time derivative of the responsibilities through the quantities $\partial_t \log \varphi_{i,t}^d$; see Lemma B.10 and Lemma B.11.

- The final condition controls the annealing contribution. Recall

$$J_{\mathrm{ann}}^d(T) := \frac{1}{4}\int_0^T \int_{\mathbb{R}^d} \left\|(\Gamma^d)^{-1/2}\frac{1}{2T}C^d\nabla \log \rho_t^d(x)\right\|^2 \rho_t^d(\,\mathrm{d}x)\,\mathrm{d}t.$$

In the diagonal Gaussian-mixture setting, this contribution is controlled by

$$\sum_{j\geq 1} \frac{\lambda_j^2}{\gamma_j \underline{\sigma}_j} < \infty. \tag{15}$$

This is the estimate proved in Lemma B.2.

Although the conditions (9)–(15) may be hard to parse without following the proof in Appendix B.1, they make explicit the two-sided role of the preconditioner discussed throughout the paper. The annealing condition (15) favors larger γ_j, while conditions (10), (11), (13), and (14) require enough tail damping of γ_j to prevent discretization errors from accumulating across high-frequency coordinates. In simple regimes, such as power-law spectral decay, this trade-off can be made explicit; see Proposition C.2, where balancing the annealing and discretization tails gives $\gamma_j \asymp \lambda_j^{2/3}$ for Gaussian mixtures with common-covariance tails.

ELP Beyond the EM Stability Restriction. The next proposition shows that, in contrast with what we saw for EM, the conditions ensuring a dimension-uniform bound for the ELP scheme do not force the initial smoothed law of ALD to remain uniformly close to the target across dimensions. The proof is given in Appendix B.2.

Proposition 4.1. *Let e_1^d be the first canonical basis vector of $\mathbb{R}^d$. Fix $a > 1$ and consider the two-component target $\rho_\star^d = \frac{1}{2}\mathcal{N}(ae_1^d, \Sigma_1^d) + \frac{1}{2}\mathcal{N}(-ae_1^d, \Sigma_2^d)$, where $\Sigma_1^d = \mathrm{Diag}(\sigma_j)_{j=1}^d$ and $\Sigma_2^d = \mathrm{Diag}(\sigma_j + \delta_j)_{j=1}^d$. Let $\sigma_j = \lambda_j = j^{-6}$ and $\gamma_j = j^{-4}$, with $\delta_1 = 0$ and $\delta_j = j^{-12}$ for $j \geq 2$. Then the summability conditions* (9)–(15) *hold, while*

$$\mathrm{KL}(\rho_\star^d \,\|\, \rho_0^d) \to \infty \qquad \text{as } d \to \infty.$$

Nevertheless, for every $T > 0$ and every mesh,

$$\sup_{d\geq 1} \mathrm{KL}\big(\rho_\star^d \,\|\, \mathrm{Law}(Y_T^d)\big) \leq \frac{1}{8T} \sum_{j\geq 1} j^{-2} + C_{\mathrm{disc}}(1+T^2)\, h_{\max} < \infty.$$

5 Numerical Experiments

We conclude with a simple numerical experiment showing the high-frequency stability issue of EM identified in Section 3, and showing that ELP remains dimension-robust even when the initial smoothed law becomes increasingly far from the target as d grows. We consider a two-component Gaussian mixture target with common covariance $\Sigma^d = \mathrm{Diag}(\sigma_j)$, $\sigma_j = \lambda_j = j^{-6}$, and preconditioner $\gamma_j = j^{-4}$, initialized from $\rho_0^d = \rho_\star^d * \mathcal{N}(0, 2SC^d)$ with $S = 5$. Then $\mathrm{KL}(\rho_\star^d \| \rho_0^d)$ grows linearly with d, while the ELP conditions (9)–(15) hold, so that the KL error can be made uniformly small in d. For EM, however, $\gamma_j/\sigma_j = j^2$, so the stability condition of Proposition 3.1 requires time steps of order d^{-2}; any fixed step size becomes unstable as d grows. Figure 1 compares the two schemes. EM rapidly becomes unstable in the high-frequency coordinates, both in KL and in the coordinate-wise variance profile at $d = 50$, whereas ELP remains close to the target scale. Further details are given in Appendix D.

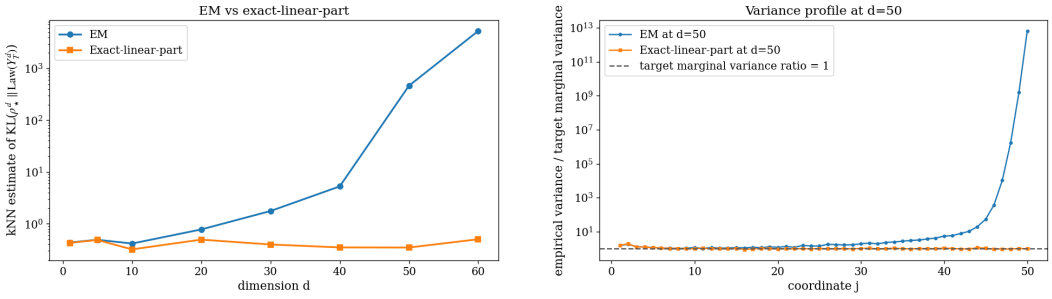

Figure 1: Left: empirical $\mathrm{KL}(\rho_\star^d \| \mathrm{Law}(Y_T^d))$ versus dimension d on a log scale. EM grows rapidly, while ELP remains stable. Right: coordinate-wise variance profile at $d = 50$, normalized by the target marginal variance. EM is unstable in the high-frequency coordinates, while ELP remains near the target scale. The KL is estimated by kNN with $k = 20$; robustness to k is reported in Appendix D.

6 Discussion and Future Work

We addressed the problem of obtaining stable annealed Langevin dynamics under successive finite-dimensional approximations of an infinite-dimensional multimodal target. To the best of our knowledge, this is the first discretization analysis of ALD with guarantees that remain uniform as the finite-dimensional approximation is refined. For Gaussian mixtures,we showed that dimension-uniform control of the KL divergence from the target to the sampler law depends both on sufficient high-frequency damping by the preconditioner and a discretization that treats the high-frequency structure of the annealed score appropriately; otherwise, additional stability conditions may arise that sharply narrow the regimes in which the sampler admits dimension-uniform KL guarantees.

We analyzed two discretization schemes. For Euler–Maruyama, we showed that treating the linear part of the annealed score by a forward Euler step imposes a stability condition which, combined with annealing-bias control, forces the smoothed law used to initialize ALD to remain uniformly close to the target across dimensions. By contrast, the ELP scheme of Section 4 achieves dimension-uniform KL control even when the initial smoothed law can be far from the target. Together with the continuous-time analysis of [2], this provides a unified picture of preconditioned ALD in infinite dimensions, covering initialization mismatch, score perturbations, and discretization error.

As is common in infinite-dimensional analysis [3, 40, 1, 2, 20], we worked under structural assumptions that make the dimension-uniform conditions explicit in terms of the relevant spectra. For example, we assumed that the mixture covariances, smoothing covariance, and preconditioner are co-diagonalizable. This allows these objects to be compared coordinate by coordinate, so that dimension-uniform control can be expressed through explicit summability conditions. Nevertheless, the model should not be viewed as simplistic: at each fixed truncation level, mixtures with diagonal covariances can still represent a broad class of multimodal distributions, especially when the number of components is allowed to vary, or even be countable. With additional operator estimates, the same strategy should extend to non-diagonal covariances and non-diagonal preconditioners, for example under a common Loewner lower-envelope assumption on the annealed component covariances.

Acknowledgments

We thank Reviewer kBG3 for the valuable discussion on OpenReview regarding our previous paper [2]. This exchange motivated us to investigate the question addressed here.

JG was supported by Agence de l'Innovation de Défense (AID) via the Centre Interdisciplinaire d'Études pour la Défense et la Sécurité (CIEDS) (projects 2021–PRODIPO and 2025–UABF). KS was supported by the Air Force Office of Scientific Research under grant FA9550261B189 and the National Science Foundation under grant DMS-2606808. MVdH acknowledges support from the Department of Energy under grant DE-SC0020345, Oxy, and the Simons Foundation under the MATH+X program.

References

[1] Lorenzo Baldassari, Josselin Garnier, Knut Sølna, and Maarten V. de Hoop. Preconditioned Langevin dynamics with score-based generative models for infinite-dimensional linear Bayesian inverse problems. In *Proceedings of the 39th International Conference on Neural Information Processing Systems*, 2025.

[2] Lorenzo Baldassari, Josselin Garnier, Knut Solna, and Maarten V de Hoop. Dimension-free multimodal sampling via preconditioned annealed langevin dynamics. *arXiv preprint arXiv:2602.01449*, 2026.

[3] Lorenzo Baldassari, Ali Siahkoohi, Josselin Garnier, Knut Sølna, and Maarten V. de Hoop. Conditional score-based diffusion models for Bayesian inference in infinite dimensions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

[4] Lorenzo Baldassari, Ali Siahkoohi, Josselin Garnier, Knut Solna, and Maarten V de Hoop. Taming score-based diffusion priors for infinite-dimensional nonlinear inverse problems. *arXiv preprint arXiv:2405.15676*, 2024.

[5] Alexandros Beskos, Mark Girolami, Shiwei Lan, Patrick E Farrell, and Andrew M Stuart. Geometric MCMC for infinite-dimensional inverse problems. *Journal of Computational Physics*, 335:327–351, 2017.

[6] Robert B Best and Gerhard Hummer. Coordinate-dependent diffusion in protein folding. *Proceedings of the National Academy of Sciences*, 107(3):1088–1093, 2010.

[7] Vladimir Igorevich Bogachev. *Gaussian Measures*. Number 62. American Mathematical Soc., 1998.

[8] Sam Bond-Taylor and Chris G Willcocks. ∞-diff: Infinite resolution diffusion with subsampled mollified states. *International Conference on Learning Representations*, 2024.

[9] Eugen Bronasco, Benedict Leimkuhler, Dominic Phillips, and Gilles Vilmart. Efficient Langevin sampling with position-dependent diffusion. *arXiv preprint arXiv:2501.02943*, 2025.

[10] Patrick Cattiaux, Paula Cordero-Encinar, and Arnaud Guillin. Diffusion annealed Langevin dynamics: a theoretical study. *arXiv preprint arXiv:2511.10406*, 2025.

[11] Omar Chehab, Anna Korba, Austin Stromme, and Adrien Vacher. Provable convergence and limitations of geometric tempering for Langevin dynamics. *International Conference on Learning Representations*, 2025.

[12] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

[13] Paula Cordero-Encinar, O Deniz Akyildiz, and Andrew B Duncan. Non-asymptotic analysis of diffusion annealed Langevin Monte Carlo for generative modelling. *arXiv preprint arXiv:2502.09306*, 2025.

[14] Simon L Cotter, Gareth O Roberts, Andrew M Stuart, and David White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, pages 424–446, 2013.

[15] Tiangang Cui, Kody JH Law, and Youssef M Marzouk. Dimension-independent likelihood-informed MCMC. *Journal of Computational Physics*, 304:109–137, 2016.

[16] Tiangang Cui, Xin Tong, and Olivier Zahm. Optimal Riemannian metric for Poincaré inequalities and how to ideally precondition Langevin dynamics. *arXiv preprint arXiv:2404.02554*, 2024.

[17] Arnak S Dalalyan and Alexandre B Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012.

[18] Jing Dong and Xin T Tong. Spectral gap of replica exchange Langevin diffusion on mixture distributions. *Stochastic Processes and their Applications*, 151:451–489, 2022.

[19] Giulio Franzese, Giulio Corallo, Simone Rossi, Markus Heinonen, Maurizio Filippone, and Pietro Michiardi. Continuous-time functional diffusion processes. *Advances in Neural Information Processing Systems*, 36:37370–37400, 2023.

[20] Giulio Franzese and Pietro Michiardi. Generative diffusion models in infinite dimensions: a survey. *Philosophical Transactions A*, 383(2299):20240322, 2025.

[21] Saul B Gelfand and Sanjoy K Mitter. On sampling methods and annealing algorithms. Technical report, 1990.

[22] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[23] Wei Guo, Molei Tao, and Yongxin Chen. Provable benefit of annealed Langevin Monte Carlo for non-log-concave sampling. *International Conference on Learning Representations*, 2025.

[24] Paul Hagemann, Sophie Mildenberger, Lars Ruthotto, Gabriele Steidl, and Nicole Tianjiao Yang. Multilevel diffusion: Infinite dimensional score-based diffusion models for image generation. *SIAM Journal on Mathematics of Data Science*, 7(3):1337–1366, 2025.

[25] Martin Hairer, Andrew M Stuart, and Sebastian J Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24:2455–2490, 2014.

[26] Gerhard Hummer. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New Journal of Physics*, 7(1):34–34, 2005.

[27] Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion generative models in infinite dimensions. *International Conference on Artificial Intelligence and Statistics*, 2023.

[28] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[29] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin, 1992.

[30] Yoshio Komori and Kevin Burrage. A stochastic exponential euler scheme for simulation of stiff biochemical reaction systems. *BIT Numerical Mathematics*, 54(4):1067–1085, 2014.

[31] Tony Lelièvre, Grigorios A Pavliotis, Geneviève Robin, Régis Santet, and Gabriel Stoltz. Optimizing the diffusion coefficient of overdamped Langevin dynamics. *arXiv preprint arXiv:2404.12087*, 2024.

[32] Tony Lelièvre, Régis Santet, and Gabriel Stoltz. Improving sampling by modifying the effective diffusion. *Journal of Computational Physics*, 541:114313, 2025.

[33] Jae Hyun Lim, Nikola B Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, et al. Score-based diffusion models in function space. *Journal of Machine Learning Research*, 26(158):1–62, 2025.

[34] Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019.

[35] Enzo Marinari and Giorgio Parisi. Simulated tempering: a new Monte Carlo scheme. *EPL (Europhysics Letters)*, 19(6):451–458, 1992.

[36] Geoffrey J McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2000.

[37] Radford M Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6(4):353–366, 1996.

[38] Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[39] Fernando Pérez-Cruz. Kullback-Leibler divergence estimation of continuous distributions. In *2008 IEEE international symposium on information theory*, pages 1666–1670. IEEE, 2008.

[40] Jakiw Pidstrigach, Youssef Marzouk, Sebastian Reich, and Sven Wang. Infinite-dimensional diffusion models. *Journal of Machine Learning Research*, 25(414):1–52, 2024.

[41] Luc Rey-Bellet and Konstantinos Spiliopoulos. Improving the convergence of reversible samplers. *Journal of Statistical Physics*, 164(3):472–494, 2016.

[42] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

[43] Gareth O Roberts and Osnat Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

[44] André Schlichting. Poincaré and log–sobolev inequalities for mixtures. *Entropy*, 21(1):89, 2019.

[45] Chunmei Shi, Yu Xiao, and Chiping Zhang. The convergence and ms stability of exponential Euler method for semilinear stochastic differential equations. *Abstract and Applied Analysis*, 2012:350407, 2012.

[46] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[47] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in Neural Information Processing Systems*, 33:12438–12448, 2020.

[48] Andrew M Stuart. Inverse problems: a Bayesian perspective. *Acta numerica*, 19:451–559, 2010.

[49] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

[50] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *International Conference on Learning Representations*, 2023.

[51] Nicolas Zilberstein, Chris Dick, Rahman Doost-Mohammady, Ashutosh Sabharwal, and Santiago Segarra. Annealed Langevin dynamics for massive MIMO detection. *IEEE Transactions on Wireless Communications*, 22(6):3762–3776, 2022.

[52] Nicolas Zilberstein, Ashutosh Sabharwal, and Santiago Segarra. Solving linear inverse problems using higher-order annealed Langevin diffusion. *IEEE Transactions on Signal Processing*, 72:492–505, 2024.

A Proofs of Section 3

A.1 Proof of Proposition 3.2

Since the two mixture components differ only in the first coordinate, the annealed path is

$$\pi_t^d = \frac{1}{2}\mathcal{N}(ae_1^d, D_t^d) + \frac{1}{2}\mathcal{N}(-ae_1^d, D_t^d), \qquad D_t^d = \Sigma^d + \kappa_t C^d.$$

Its score is explicit:

$$\nabla \log \pi_t^d(x) = \left(-\frac{x_1}{d_1(t)} + \frac{a}{d_1(t)} \tanh\left(\frac{ax_1}{d_1(t)}\right), -\frac{x_2}{d_2(t)}, \dots, -\frac{x_d}{d_d(t)} \right),$$

where

$$d_j(t) := \sigma_j + \kappa_t \lambda_j.$$

Hence the ALD dynamics decouples into a nonlinear first coordinate and Gaussian tail coordinates. More precisely, for every $j \geq 2$,

$$dX_{t,j} = -\frac{\gamma_j}{d_j(t)} X_{t,j}\, dt + \sqrt{2\gamma_j}\, dW_{t,j}, \qquad X_{0,j} \sim \mathcal{N}(0, \sigma_j + \lambda_j).$$

It follows that the target and the terminal law factorize as

$$\pi_\star^d = \nu_{\star,1} \otimes \bigotimes_{j=2}^{d} \mathcal{N}(0, \sigma_j), \qquad \mathrm{Law}(X_T^d) = \mu_{T,1} \otimes \bigotimes_{j=2}^{d} \mathcal{N}(0, p_j(T)),$$

for suitable one-dimensional laws $\nu_{\star,1}$ and $\mu_{T,1}$, where $p_j(T)$ is the variance of the j-th coordinate at time T. By additivity of relative entropy,

$$\mathrm{KL}(\pi_\star^d \,\|\, \mathrm{Law}(X_T^d)) = \mathrm{KL}(\nu_{\star,1} \,\|\, \mu_{T,1}) + \sum_{j=2}^{d} \mathrm{KL}\big(\mathcal{N}(0, \sigma_j) \,\|\, \mathcal{N}(0, p_j(T))\big). \tag{16}$$

We now compute the tail terms explicitly. For $j \geq 2$, the variance $p_j(t)$ solves

$$p_j'(t) = 2\gamma_j \left(1 - \frac{p_j(t)}{d_j(t)}\right), \qquad p_j(0) = \sigma_j + \lambda_j.$$

Set

$$r_j := \frac{\lambda_j}{\sigma_j}, \qquad \alpha_j := \frac{2\gamma_j T}{\lambda_j}, \qquad \Psi(\alpha, r) := \int_1^{1+r} u^{-\alpha}\, du.$$

A direct computation gives

$$\frac{p_j(T) - \sigma_j}{\sigma_j} = \Psi(\alpha_j, r_j).$$

Therefore

$$\mathrm{KL}\big(\mathcal{N}(0, \sigma_j) \,\|\, \mathcal{N}(0, p_j(T))\big) = F\big(\Psi(\alpha_j, r_j)\big),$$

where

$$F(u) := \frac{1}{2}\left(\log(1+u) - \frac{u}{1+u}\right), \qquad u \geq 0.$$

Now let

$$M := \sup_{j \geq 2} \frac{\gamma_j}{\sigma_j} < \infty.$$

Then

$$\alpha_j = \frac{2\gamma_j T}{\lambda_j} = \frac{2T}{r_j}\frac{\gamma_j}{\sigma_j} \leq \frac{2TM}{r_j}.$$

Since $\Psi(\alpha, r)$ is decreasing in α, writing $c := 2TM$ yields

$$\Psi(\alpha_j, r_j) \geq \Psi(c/r_j, r_j).$$

Moreover,

$$\Psi(c/r, r) = r \int_0^1 (1 + rs)^{-c/r}\, ds.$$

Using $\log(1 + rs) \le rs$, we obtain

$$(1 + rs)^{-c/r} = \exp\Big(-\frac{c}{r}\log(1 + rs)\Big) \ge e^{-cs},$$

and therefore

$$\Psi(c/r, r) \ge r \int_0^1 e^{-cs}\, ds = \beta r, \qquad \beta := \frac{1 - e^{-c}}{c} > 0.$$

Thus

$$\Psi(\alpha_j, r_j) \ge \beta r_j \qquad \text{for all } j \ge 2.$$

Combining this with (16), the assumed uniform bound on the true annealing bias implies

$$\sup_{d \ge 1} \sum_{j=2}^{d} F(\beta r_j) < \infty.$$

If infinitely many r_j were larger than $1/\beta$, then infinitely many summands would be bounded below by the positive constant $F(1)$, which is impossible. Hence $r_j \le 1/\beta$ for all but finitely many j. For such j, we have $\beta r_j \le 1$, and since

$$F(u) = \int_0^u \frac{s}{2(1 + s)^2}\, ds,$$

it follows that for $0 \le u \le 1$,

$$F(u) \ge \frac{u^2}{16}.$$

Hence, for all sufficiently large j,

$$F(\beta r_j) \ge \frac{\beta^2}{16} r_j^2.$$

Therefore

$$\sum_{j \ge 2} r_j^2 < \infty.$$

Finally, the initial annealed law also factorizes:

$$\pi_0^d = \nu_{0,1} \otimes \bigotimes_{j=2}^{d} \mathcal{N}(0, \sigma_j + \lambda_j),$$

for a suitable one-dimensional law $\nu_{0,1}$. Thus

$$\mathrm{KL}(\pi_\star^d \,\|\, \pi_0^d) = \mathrm{KL}(\nu_{\star,1} \,\|\, \nu_{0,1}) + \sum_{j=2}^{d} F(r_j).$$

Since $\sum_{j \ge 2} r_j^2 < \infty$, we have $r_j \to 0$, so for all sufficiently large j also $r_j \le 1$, and then

$$F(r_j) \le \frac{r_j^2}{4}.$$

Hence

$$\sum_{j \ge 2} F(r_j) < \infty,$$

and the first-coordinate contribution is a fixed constant independent of d. Therefore

$$\sup_{d \ge 1} \mathrm{KL}(\pi_\star^d \,\|\, \pi_0^d) < \infty.$$

This proves the claim.

A.2 Proof of Proposition 3.3

By joint convexity of relative entropy,

$$\mathrm{KL}(\rho_\star^d \,\|\, \rho_0^d) \le \sum_{i\in I} w_i \,\mathrm{KL}\Big(\mathcal{N}(m_i^d, \Sigma_i^d) \,\Big\|\, \mathcal{N}(m_i^d, \Sigma_i^d + C^d)\Big).$$

Since the means coincide, the Gaussian relative entropy formula yields

$$\mathrm{KL}(\rho_\star^d \,\|\, \rho_0^d) \le \frac{1}{2}\sum_{i\in I} w_i \sum_{j=1}^{d}\left[\log\left(1+\frac{\lambda_j}{\sigma_{ij}}\right) - \frac{\lambda_j}{\sigma_{ij}+\lambda_j}\right].$$

For $r \ge 0$, the scalar function

$$f(r) := \log(1+r) - \frac{r}{1+r}$$

satisfies $f(r) \le r^2/2$. Therefore

$$\mathrm{KL}(\rho_\star^d \,\|\, \rho_0^d) \le \frac{1}{4}\sum_{i\in I} w_i \sum_{j=1}^{d}\frac{\lambda_j^2}{\sigma_{ij}^2}.$$

B Proofs of Section 4

B.1 Proof of Theorem 4.1

The proof has two parts. Recall that the score is decomposed as

$$\nabla \log \rho_t^d(x) = -(B_t^d)^{-1}x + G_t^d(x),$$

where G_t^d is the nonlinear mixture correction. The ELP interpolation Y^d integrates the stiff diagonal linear part exactly, but freezes this nonlinear correction on each discretization interval $[t_n, t_{n+1})$ at its left-endpoint value $G_{t_n}^d(Y_{t_n}^d)$. To compare this scheme with the target, we introduce an auxiliary process $\widehat{X}^d$ whose marginals exactly follow the annealing path, that is, $\mathrm{Law}(\widehat{X}_t^d) = \rho_t^d$. When the two dynamics are compared along the auxiliary path, the auxiliary process uses the current nonlinear correction $G_t^d(\widehat{X}_t^d)$, whereas the ELP interpolation uses the frozen correction $G_{t_n}^d(\widehat{X}_{t_n}^d)$. Thus the path-space KL estimate over $[0,T]$ between the law of the auxiliary process and the law of the ELP interpolation separates the error into an *annealing contribution*, coming from the transport velocity of the annealing path, and a *freezing defect*

$$G_t^d(\widehat{X}_t^d) - G_{t_n}^d(\widehat{X}_{t_n}^d), \qquad t \in [t_n, t_{n+1}).$$

The second part of the proof verifies that the coefficient assumptions (9)–(15) control both contributions uniformly in d, while tracking the dependence on the horizon T.

In this proof, we shall split the freezing defect into a *spatial part*, where the time is fixed but the spatial argument changes, and a *temporal part*, where the spatial argument is fixed but the time changes:

$$G_t^d(\widehat{X}_t^d) - G_{t_n}^d(\widehat{X}_{t_n}^d) = \big(G_t^d(\widehat{X}_t^d) - G_t^d(\widehat{X}_{t_n}^d)\big) + \big(G_t^d(\widehat{X}_{t_n}^d) - G_{t_n}^d(\widehat{X}_{t_n}^d)\big).$$

The estimates below are organized around this decomposition.

In what follows, for notational simplicity, we do not distinguish between finite and countable mixtures. In the countable case, the argument can be justified by applying the estimates to finite partial mixtures and then passing to the limit. The normalization constants of the partial mixtures converge to one and cancel in the responsibility formulas; moreover, the constants in the estimates depend only on the uniform quantities appearing in the assumptions.

Coefficient Notation. We start by recording the coefficient notation used throughout. For each $i \in I$ and $j \ge 1$, we write

$$v_{i,t,j} := \sigma_{ij} + \kappa_t \lambda_j, \qquad D_{i,t}^d = \mathrm{Diag}(v_{i,t,1}, \dots, v_{i,t,d}).$$

We also define

$$\delta\sigma_j := \overline{\sigma}_j - \underline{\sigma}_j, \qquad \Delta m_j := \sup_{i,\ell\in I} |m_{ij} - m_{\ell j}|,$$

and

$$A_j := \frac{\delta\sigma_j}{\underline{\sigma}_j^2}, \qquad B_j := \frac{\Delta m_j}{\underline{\sigma}_j} + \frac{\overline{m}_j\,\delta\sigma_j}{\underline{\sigma}_j^2}, \qquad D_j := \frac{\lambda_j\delta\sigma_j}{\underline{\sigma}_j^3}.$$

With this notation, the summability assumptions (9)–(15) become

$$\sum_{j\geq 1}(\overline{\sigma}_j + \lambda_j) < \infty, \qquad \sum_{j\geq 1}\overline{m}_j^2 < \infty,$$

$$\sum_{j\geq 1}\gamma_j < \infty, \qquad \sum_{j\geq 1}\gamma_j^2\frac{\overline{\sigma}_j + \lambda_j}{\underline{\sigma}_j^2} < \infty, \qquad \sum_{j\geq 1}\gamma_j^2\frac{\overline{m}_j^2}{\underline{\sigma}_j^2} < \infty,$$

$$\sum_{j\geq 1}\gamma_j A_j^2 < \infty, \qquad \sum_{j\geq 1}\gamma_j\frac{\overline{m}_j^2}{\underline{\sigma}_j^2} < \infty,$$

$$\sum_{j\geq 1}A_j^2(\overline{\sigma}_j + \lambda_j + \overline{m}_j^2) < \infty, \qquad \sum_{j\geq 1}B_j^2 < \infty,$$

$$\sum_{j\geq 1}D_j(\overline{\sigma}_j + \lambda_j + \overline{m}_j^2) < \infty, \qquad \sum_{j\geq 1}\gamma_j D_j^2(\overline{\sigma}_j + \lambda_j + \overline{m}_j^2) < \infty,$$

$$\sum_{j\geq 1}\lambda_j^2\frac{B_j^2}{\underline{\sigma}_j^2} < \infty, \qquad \sum_{j\geq 1}\gamma_j\lambda_j^2\frac{\overline{m}_j^2}{\underline{\sigma}_j^4} < \infty,$$

and

$$\sum_{j\geq 1}\frac{\lambda_j^2}{\gamma_j\underline{\sigma}_j} < \infty.$$

Recall that

$$G_t^d(x) = \sum_{i\in I} p_{i,t}^d(x)c_{i,t}^d(x),$$

where $p_{i,t}^d(x)$ are the mixture responsibilities and

$$c_{i,t}^d(x) = \Big((B_t^d)^{-1} - (D_{i,t}^d)^{-1}\Big)x + (D_{i,t}^d)^{-1}m_i^d$$

is the affine correction associated with the i-th component. The terms involving A_j control the spatial variation of these affine corrections; the terms involving B_j control the spatial variation of the responsibilities; and the terms involving D_j control the time variation caused by the annealing schedule. The final condition involving $\lambda_j^2/(\gamma_j\underline{\sigma}_j)$ controls the annealing energy.

The Auxiliary Path and the Quantities to Be Estimated. On each discretization interval $[t_n, t_{n+1})$, the ELP scheme freezes G_t^d at time t_n, and therefore does not follow the annealing path exactly. To compare it with the target, we introduce a reference process whose marginals are exactly ρ_t^d. The extra drift below is the transport velocity associated with the evolution of the annealing path:

$$u_t^d(x) := \frac{1}{2T}C^d\nabla\log\rho_t^d(x). \tag{17}$$

Let $(\widehat{X}_t^d)_{t\in[0,T]}$ solve

$$d\widehat{X}_t^d = -\Gamma^d(B_t^d)^{-1}\widehat{X}_t^d\,dt + \Gamma^d G_t^d(\widehat{X}_t^d)\,dt + u_t^d(\widehat{X}_t^d)\,dt + \sqrt{2\Gamma^d}\,dW_t^d, \qquad \widehat{X}_0^d \sim \rho_0^d. \tag{18}$$

Since

$$\partial_t\rho_t^d = -\nabla\cdot\left(\frac{1}{2T}C^d\nabla\log\rho_t^d\,\rho_t^d\right),$$

and the Langevin part with drift $\Gamma^d\nabla\log\rho_t^d$ leaves ρ_t^d instantaneously invariant, the Fokker-Planck equation associated with (18) is solved by ρ_t^d. Hence

$$\mathrm{Law}(\widehat{X}_t^d) = \rho_t^d, \qquad t\in[0,T],$$

and in particular

$$\mathrm{Law}(\widehat{X}_T^d) = \rho_T^d = \rho_\star^d.$$

We shall use the corresponding path-matching energy

$$J_{\mathrm{ann}}^d(T) := \frac{1}{4}\int_0^T \int_{\mathbb{R}^d} \left\| (\Gamma^d)^{-1/2} u_t^d(x) \right\|^2 \rho_t^d(dx)\, dt. \tag{19}$$

Following [23], which builds on [17, 12], the comparison with the ELP interpolation is based on a Girsanov estimate on path space over $[0,T]$. Consequently, the relevant quantity is the drift mismatch between the auxiliary process and the ELP interpolation. On an interval $[t_n, t_{n+1})$, this mismatch is

$$u_t^d(\widehat{X}_t^d) + \Gamma^d\Big(G_t^d(\widehat{X}_t^d) - G_{t_n}^d(\widehat{X}_{t_n}^d)\Big).$$

After weighting by $(\Gamma^d)^{-1/2}$, the two quantities to control are therefore

$$(\Gamma^d)^{-1/2} u_t^d(\widehat{X}_t^d)$$

and

$$\Delta_t^d := (\Gamma^d)^{1/2}\Big(G_t^d(\widehat{X}_t^d) - G_{t_n}^d(\widehat{X}_{t_n}^d)\Big). \tag{20}$$

The first term gives the *annealing contribution* $J_{\mathrm{ann}}^d(T)$. The second term is the *freezing defect*. As anticipated, we decompose it as

$$\Delta_t^d = \underbrace{(\Gamma^d)^{1/2}\big(G_t^d(\widehat{X}_t^d) - G_t^d(\widehat{X}_{t_n}^d)\big)}_{\text{spatial freezing defect}} + \underbrace{(\Gamma^d)^{1/2}\big(G_t^d(\widehat{X}_{t_n}^d) - G_{t_n}^d(\widehat{X}_{t_n}^d)\big)}_{\text{temporal freezing defect}}.$$

The proof will reduce to three estimates:

$$J_{\mathrm{ann}}^d(T) \lesssim \frac{1}{T}\sum_{j\ge 1} \frac{\lambda_j^2}{\gamma_j \underline{\sigma}_j},$$

$$\mathbb{E}\left\| (\Gamma^d)^{1/2}\big(G_t^d(\widehat{X}_t^d) - G_t^d(\widehat{X}_s^d)\big)\right\|^2 \lesssim \left(1 + T + \frac{1}{T}\right)(t-s),$$

and

$$\mathbb{E}\left\| (\Gamma^d)^{1/2}\big(G_t^d(\widehat{X}_s^d) - G_s^d(\widehat{X}_s^d)\big)\right\|^2 \lesssim \frac{(t-s)^2}{T^2}.$$

The lemmas below prove these three estimates.

Moment and Transport Estimates. We begin with two preliminary estimates. The first gives uniform moment bounds for the annealing path ρ_t^d, and hence for the marginals of $\widehat{X}_t^d$. These bounds are needed later because the freezing defect contains the random fields

$$G_t^d(\widehat{X}_t^d) \qquad \text{and} \qquad G_t^d(\widehat{X}_{t_n}^d),$$

and because the spatial freezing defect will require control of increments

$$\widehat{X}_t^d - \widehat{X}_s^d.$$

The second estimate controls the transport field u_t^d. It yields the bound on the annealing contribution $J_{\mathrm{ann}}^d(T)$ and also provides a fourth-moment estimate for $u_t^d(\widehat{X}_t^d)$, which will enter the increment estimates below.

Lemma B.1. *Under* (9)*, there exist constants M_2, M_4, M_8, independent of d, t, and T, such that*

$$\sup_{d\ge 1}\sup_{t\in[0,T]} \int_{\mathbb{R}^d} \|x\|^q\, \rho_t^d(dx) \le M_q, \qquad q = 2, 4, 8.$$

Proof. Fix i, d, t. Since

$$v_{i,t,j} = \sigma_{ij} + \kappa_t \lambda_j \le \overline{\sigma}_j + \lambda_j,$$

we have

$$\operatorname{Tr}(D_{i,t}^d) \le \sum_{j\ge 1} (\overline{\sigma}_j + \lambda_j) < \infty.$$

Moreover,

$$\|m_i^d\|^2 = \sum_{j=1}^d m_{ij}^2 \le \sum_{j\ge 1} \overline{m}_j^2 < \infty.$$

Thus the second moments are uniformly bounded over components, d, and t.

For the fourth moment, let $Z = m + \xi$, with $\xi \sim \mathcal{N}(0, D)$ and D diagonal. Then

$$\|Z\|^4 \le 8\|m\|^4 + 8\|\xi\|^4.$$

Now

$$\mathbb{E}\|\xi\|^4 = \mathbb{E}\left(\sum_{j=1}^d \xi_j^2\right)^2 \le 3\left(\sum_{j=1}^d D_{jj}\right)^2 = 3(\operatorname{Tr} D)^2.$$

Hence

$$\mathbb{E}\|Z\|^4 \le 8\|m\|^4 + 24(\operatorname{Tr} D)^2.$$

For the eighth moment, again

$$\|Z\|^8 \le 2^7\|m\|^8 + 2^7\|\xi\|^8.$$

By Minkowski's inequality in L^4,

$$\left(\mathbb{E}\|\xi\|^8\right)^{1/4} = \left\|\sum_{j=1}^d \xi_j^2\right\|_{L^4} \le \sum_{j=1}^d \|\xi_j^2\|_{L^4} = \sum_{j=1}^d \left(\mathbb{E}|\xi_j|^8\right)^{1/4}.$$

For a centered one-dimensional Gaussian with variance D_{jj},

$$\mathbb{E}|\xi_j|^8 = 105 D_{jj}^4.$$

Therefore

$$\mathbb{E}\|\xi\|^8 \le 105(\operatorname{Tr} D)^4.$$

It follows that

$$\mathbb{E}\|Z\|^8 \le 2^7\|m\|^8 + 2^7 \cdot 105(\operatorname{Tr} D)^4.$$

Applying these estimates to each Gaussian component $\mathcal{N}(m_i^d, D_{i,t}^d)$ and using the uniform bounds on $\left\|m_i^d\right\|^2$ and $\operatorname{Tr} D_{i,t}^d$ proves the required uniform fourth and eighth moment estimates. □

We next control the transport field associated with the annealing path. This is the term that produces the annealing contribution in the path-space KL estimate.

Lemma B.2. *Under $\sum_j \gamma_j < \infty$ and* (15)*, we have*

$$\sup_{d\ge 1} J_{\mathrm{ann}}^d(T) \le \frac{1}{16T} \sum_{j\ge 1} \frac{\lambda_j^2}{\gamma_j \underline{\sigma}_j}.$$

Moreover, there exists a constant C_u, independent of d and T, such that

$$\sup_{d\ge 1} \sup_{t\in[0,T]} \mathbb{E}\|u_t^d(\widehat{X}_t^d)\|^4 \le \frac{C_u}{T^4}.$$

Proof. The argument follows that of [2, Theorem 3.1]. Write

$$s_{i,t}^d(x) := -(D_{i,t}^d)^{-1}(x - m_i^d).$$

Then

$$\nabla \log \rho_t^d(x) = \sum_{i\in I} p_{i,t}^d(x) s_{i,t}^d(x).$$

By Jensen's inequality,

$$\left\|(\Gamma^d)^{-1/2}u_t^d(x)\right\|^2 \le \frac{1}{4T^2}\sum_{i\in I} p_{i,t}^d(x)\left\|(\Gamma^d)^{-1/2}C^d s_{i,t}^d(x)\right\|^2.$$

Integrating with respect to ρ_t^d and using the mixture representation,

$$\int_{\mathbb{R}^d}\left\|(\Gamma^d)^{-1/2}u_t^d(x)\right\|^2 \rho_t^d(dx) \le \frac{1}{4T^2}\sum_{i\in I} w_i \sum_{j=1}^d \frac{\lambda_j^2}{\gamma_j}\mathbb{E}_{\mathcal{N}(m_i^d, D_{i,t}^d)}\left[\frac{(X_j - m_{ij})^2}{v_{i,t,j}^2}\right].$$

Since

$$\mathbb{E}_{\mathcal{N}(m_i^d, D_{i,t}^d)}\left[(X_j - m_{ij})^2\right] = v_{i,t,j},$$

we get

$$\int_{\mathbb{R}^d}\left\|(\Gamma^d)^{-1/2}u_t^d(x)\right\|^2 \rho_t^d(dx) \le \frac{1}{4T^2}\sum_{j=1}^d \frac{\lambda_j^2}{\gamma_j \underline{\sigma}_j}.$$

Integrating in time and multiplying by $1/4$ according to (19) gives the desired bound on $J_{\mathrm{ann}}^d(T)$.

For the fourth moment, Jensen's inequality gives

$$\|u_t^d(x)\|^4 \le \frac{1}{16T^4}\sum_{i\in I} p_{i,t}^d(x)\|C^d s_{i,t}^d(x)\|^4.$$

Integrating against ρ_t^d and using the mixture decomposition,

$$\mathbb{E}\|u_t^d(\widehat{X}_t^d)\|^4 \le \frac{1}{16T^4}\sum_{i\in I} w_i \mathbb{E}_{\mathcal{N}(m_i^d, D_{i,t}^d)}\|C^d s_{i,t}^d(X)\|^4.$$

Under the i-th component $\mathcal{N}(m_i^d, D_{i,t}^d)$, the coordinates of $C^d s_{i,t}^d(X)$ are independent centered Gaussians with variances $\lambda_j^2/v_{i,t,j}$. Hence

$$\mathbb{E}_{\mathcal{N}(m_i^d, D_{i,t}^d)}\|C^d s_{i,t}^d(X)\|^4 = \Big(\sum_{j=1}^d \frac{\lambda_j^2}{v_{i,t,j}}\Big)^2 + 2\sum_{j=1}^d \frac{\lambda_j^4}{v_{i,t,j}^2} \le 3\left(\sum_{j=1}^d \frac{\lambda_j^2}{v_{i,t,j}}\right)^2 \le 3\left(\sum_{j\ge 1}\frac{\lambda_j^2}{\underline{\sigma}_j}\right)^2.$$

Under (15),

$$\sum_{j\ge 1}\frac{\lambda_j^2}{\underline{\sigma}_j} \le \left(\sup_{j\ge 1}\gamma_j\right)\sum_{j\ge 1}\frac{\lambda_j^2}{\gamma_j\underline{\sigma}_j} < \infty,$$

because $\sum_j \gamma_j < \infty$ implies $\sup_j \gamma_j < \infty$. Thus the fourth-moment estimate holds with

$$C_u := \frac{3}{16}\left(\sum_{j\ge 1}\frac{\lambda_j^2}{\underline{\sigma}_j}\right)^2,$$

which is independent of d and T. □

Drift and Increment Estimates. The next estimates are used to control the spatial freezing defect. We will compare

$$G_t^d(\widehat{X}_t^d) \qquad \text{and} \qquad G_t^d(\widehat{X}_s^d).$$

The spatial variation of G_t^d is controlled by a Lipschitz factor with polynomial growth:

$$\|(\Gamma^d)^{1/2}(G_t^d(x) - G_t^d(y))\| \lesssim L_t^d(x,y)\|x - y\|.$$

After substituting $x = \widehat{X}_t^d$ and $y = \widehat{X}_s^d$, the spatial freezing estimate therefore involves a product of the form

$$\mathbb{E}\Big[L_t^d(\widehat{X}_t^d, \widehat{X}_s^d)^2\,\|\widehat{X}_t^d - \widehat{X}_s^d\|^2\Big].$$

The moment bounds for the annealing path control the first factor, while Hölder's inequality leaves an L^4 norm of the increment. This is why we need a fourth-moment estimate for the auxiliary path:

$$\mathbb{E}\|\widehat{X}_t^d - \widehat{X}_s^d\|^4 \le C\left(1 + T^2 + \frac{1}{T^2}\right)(t-s)^2.$$

To obtain this increment estimate, we use the SDE for the auxiliary process. Writing its drift as

$$b_t^d(x) = -\Gamma^d (B_t^d)^{-1} x + \Gamma^d G_t^d(x) + u_t^d(x),$$

we have

$$\widehat{X}_t^d - \widehat{X}_s^d = \int_s^t b_r^d(\widehat{X}_r^d)\, dr + \sqrt{2\Gamma^d}(W_t^d - W_s^d).$$

Therefore the required fourth-moment increment bound follows from two contributions: a fourth-moment bound on the drift $b_t^d(\widehat{X}_t^d)$ and the standard fourth-moment bound for the Brownian increment. The lemmas below establish these contributions in order: first the linear part of the drift, then the nonlinear correction, then the full auxiliary drift, and finally the increment estimate.

We begin with the diagonal linear part of the drift.

Lemma B.3. *Under* (10)*, there exists a constant $B_{\mathrm{lin},4}$, independent of d and T, such that*

$$\sup_{d\ge 1}\sup_{t\in[0,T]} \mathbb{E}\left\|\Gamma^d (B_t^d)^{-1}\widehat{X}_t^d\right\|^4 \le B_{\mathrm{lin},4}.$$

Proof. Let

$$A_t^d := \Gamma^d (B_t^d)^{-1} = \mathrm{Diag}\left(\frac{\gamma_1}{b_{t,1}}, \dots, \frac{\gamma_d}{b_{t,d}}\right).$$

Under the component $\mathcal{N}(m_i^d, D_{i,t}^d)$, the random vector $A_t^d X$ is Gaussian with mean $A_t^d m_i^d$ and diagonal covariance with entries

$$\frac{\gamma_j^2}{b_{t,j}^2} v_{i,t,j}.$$

Since $b_{t,j} \ge \underline{\sigma}_j$ and $v_{i,t,j} \le \overline{\sigma}_j + \lambda_j$, we have

$$\|A_t^d m_i^d\|^2 \le \sum_{j\ge 1} \gamma_j^2 \frac{\overline{m}_j^2}{\underline{\sigma}_j^2} < \infty$$

and

$$\mathrm{Tr}\left[\mathrm{Cov}(A_t^d X)\right] \le \sum_{j\ge 1} \gamma_j^2 \frac{\overline{\sigma}_j + \lambda_j}{\underline{\sigma}_j^2} < \infty.$$

Using the Gaussian fourth-moment estimate from Lemma B.1 and averaging over the mixture proves the result. □

Next we record the coordinatewise bounds on the affine corrections $c_{i,t}^d$, which will be used both for the growth of G_t^d and later for the temporal freezing defect.

Lemma B.4. *For every i, j,*

$$c_{i,t,j}^d(x) = \alpha_{ij}(t) x_j + \beta_{ij}(t),$$

where

$$\alpha_{ij}(t) := \frac{\sigma_{ij} - \underline{\sigma}_j}{b_{t,j} v_{i,t,j}}, \qquad \beta_{ij}(t) := \frac{m_{ij}}{v_{i,t,j}}.$$

Moreover,

$$0 \le \alpha_{ij}(t) \le A_j, \qquad |\beta_{ij}(t)| \le \frac{\overline{m}_j}{\underline{\sigma}_j}.$$

Finally,

$$|\dot{\alpha}_{ij}(t)| \le \frac{2D_j}{T}, \qquad |\dot{\beta}_{ij}(t)| \le \frac{\lambda_j \overline{m}_j}{T\underline{\sigma}_j^2}.$$

Proof. The formulas for α_{ij} and β_{ij} follow directly from the definition of $c_{i,t}^d$. Since $b_{t,j}, v_{i,t,j} \geq \underline{\sigma}_j$,

$$0 \leq \alpha_{ij}(t) = \frac{\sigma_{ij} - \underline{\sigma}_j}{b_{t,j} v_{i,t,j}} \leq \frac{\delta \sigma_j}{\underline{\sigma}_j^2} = A_j, \qquad |\beta_{ij}(t)| \leq \frac{\overline{m}_j}{\underline{\sigma}_j}.$$

Differentiating and using

$$\dot{b}_{t,j} = \dot{v}_{i,t,j} = -\frac{\lambda_j}{T}$$

yields

$$\dot{\alpha}_{ij}(t) = \frac{\lambda_j(\sigma_{ij} - \underline{\sigma}_j)}{T} \left(\frac{1}{b_{t,j}^2 v_{i,t,j}} + \frac{1}{b_{t,j} v_{i,t,j}^2} \right),$$

and hence

$$|\dot{\alpha}_{ij}(t)| \leq \frac{2\lambda_j \delta \sigma_j}{T \underline{\sigma}_j^3} = \frac{2D_j}{T}.$$

Similarly,

$$\dot{\beta}_{ij}(t) = \frac{\lambda_j m_{ij}}{T v_{i,t,j}^2},$$

which gives the desired bound. □

These coordinatewise bounds imply the following weighted growth estimate for the nonlinear correction G_t^d.

Lemma B.5. *Under* (11)*, there exist constants G_0, G_1, independent of d and T, such that for every d, t, x,*

$$\|(\Gamma^d)^{1/2} G_t^d(x)\| \leq G_0 + G_1 \|x\|.$$

Proof. Set

$$C_{A,\gamma} := \sum_{j \geq 1} \gamma_j A_j^2, \qquad C_{m,\gamma} := \sum_{j \geq 1} \gamma_j \frac{\overline{m}_j^2}{\underline{\sigma}_j^2},$$

which are finite under (11). Using convexity of the norm and $G_t^d = \sum_i p_{i,t}^d c_{i,t}^d$,

$$\|(\Gamma^d)^{1/2} G_t^d(x)\| \leq \sum_{i \in I} p_{i,t}^d(x) \|(\Gamma^d)^{1/2} c_{i,t}^d(x)\|.$$

By Lemma B.4,

$$\|(\Gamma^d)^{1/2} c_{i,t}^d(x)\|^2 \leq 2 \sum_{j=1}^d \gamma_j A_j^2 x_j^2 + 2 \sum_{j=1}^d \gamma_j \frac{\overline{m}_j^2}{\underline{\sigma}_j^2}.$$

Therefore

$$\|(\Gamma^d)^{1/2} c_{i,t}^d(x)\|^2 \leq 2 C_{A,\gamma} \|x\|^2 + 2 C_{m,\gamma}.$$

Taking square roots gives

$$\|(\Gamma^d)^{1/2} c_{i,t}^d(x)\| \leq \sqrt{2 C_{m,\gamma}} + \sqrt{2 C_{A,\gamma}} \|x\|.$$

□

Combining the linear drift, the nonlinear correction, and the transport drift u_t^d gives the required fourth-moment bound for the full auxiliary drift.

Lemma B.6. *Define*

$$b_t^d(x) := -\Gamma^d (B_t^d)^{-1} x + \Gamma^d G_t^d(x) + u_t^d(x).$$

Under (9)*,* (10)*,* (11)*, and* (15)*, there exist finite constants $B_{\mathrm{dr},4}^{(0)}$ and $B_{\mathrm{dr},4}^{(1)}$, independent of d and T, such that*

$$\sup_{d \geq 1} \sup_{t \in [0,T]} \mathbb{E} \|b_t^d(\widehat{X}_t^d)\|^4 \leq B_{\mathrm{dr},4}(T), \qquad B_{\mathrm{dr},4}(T) := B_{\mathrm{dr},4}^{(0)} + \frac{B_{\mathrm{dr},4}^{(1)}}{T^4}.$$

Proof. By Lemma B.3,

$$\sup_{d,t} \mathbb{E}\|\Gamma^d (B_t^d)^{-1} \widehat{X}_t^d\|^4 \le B_{\mathrm{lin},4}.$$

Since $\sum_j \gamma_j < \infty$, also $\gamma_\infty := \sup_j \gamma_j < \infty$. Using Lemma B.5,

$$\|\Gamma^d G_t^d(x)\|^2 = \sum_{j=1}^d \gamma_j^2 G_{t,j}^d(x)^2 \le \gamma_\infty \sum_{j=1}^d \gamma_j G_{t,j}^d(x)^2 = \gamma_\infty \|(\Gamma^d)^{1/2} G_t^d(x)\|^2.$$

Therefore

$$\|\Gamma^d G_t^d(x)\|^4 \le \gamma_\infty^2 (G_0 + G_1\|x\|)^4 \le 8\gamma_\infty^2 (G_0^4 + G_1^4 \|x\|^4).$$

Taking expectation under $\mathrm{Law}(\widehat{X}_t^d) = \rho_t^d$ and using Lemma B.1 gives

$$\sup_{d,t} \mathbb{E}\|\Gamma^d G_t^d(\widehat{X}_t^d)\|^4 \le 8\gamma_\infty^2 (G_0^4 + G_1^4 M_4),$$

which is independent of T. By Lemma B.2,

$$\sup_{d,t} \mathbb{E}\|u_t^d(\widehat{X}_t^d)\|^4 \le \frac{C_u}{T^4}.$$

The result follows from

$$\|a+b+c\|^4 \le 27(\|a\|^4 + \|b\|^4 + \|c\|^4),$$

with

$$B_{\mathrm{dr},4}^{(0)} := 27\left[B_{\mathrm{lin},4} + 8\gamma_\infty^2 (G_0^4 + G_1^4 M_4)\right], \qquad B_{\mathrm{dr},4}^{(1)} := 27 C_u.$$

□

We can now turn the drift bound into the desired increment estimate.

Lemma B.7. *Under the assumptions of Lemma B.6, there exists a constant $C_{\mathrm{inc},4}$, independent of d and T, such that*

$$\mathbb{E}\|\widehat{X}_t^d - \widehat{X}_s^d\|^4 \le C_{\mathrm{inc},4}\left(1 + T^2 + \frac{1}{T^2}\right)(t-s)^2, \qquad 0 \le s \le t \le T.$$

Proof. From (18),

$$\widehat{X}_t^d - \widehat{X}_s^d = \int_s^t b_r^d(\widehat{X}_r^d)\,dr + \sqrt{2\Gamma^d}(W_t^d - W_s^d).$$

Thus

$$\mathbb{E}\|\widehat{X}_t^d - \widehat{X}_s^d\|^4 \le 8\mathbb{E}\left\|\int_s^t b_r^d(\widehat{X}_r^d)\,dr\right\|^4 + 8\mathbb{E}\|\sqrt{2\Gamma^d}(W_t^d - W_s^d)\|^4.$$

By Jensen's inequality,

$$\left\|\int_s^t z_r\,dr\right\|^4 \le (t-s)^3 \int_s^t \|z_r\|^4\,dr.$$

Using Lemma B.6, we obtain

$$\mathbb{E}\left\|\int_s^t b_r^d(\widehat{X}_r^d)\,dr\right\|^4 \le \left(B_{\mathrm{dr},4}^{(0)} + \frac{B_{\mathrm{dr},4}^{(1)}}{T^4}\right)(t-s)^4.$$

Since $t - s \le T$,

$$\left(B_{\mathrm{dr},4}^{(0)} + \frac{B_{\mathrm{dr},4}^{(1)}}{T^4}\right)(t-s)^4 \le \left(B_{\mathrm{dr},4}^{(0)} T^2 + \frac{B_{\mathrm{dr},4}^{(1)}}{T^2}\right)(t-s)^2.$$

The Brownian increment is centered Gaussian with covariance $2(t-s)\Gamma^d$. Hence

$$\mathbb{E}\|\sqrt{2\Gamma^d}(W_t^d - W_s^d)\|^4 = 4(t-s)^2\left[\mathrm{Tr}(\Gamma^d)^2 + 2\,\mathrm{Tr}((\Gamma^d)^2)\right].$$

Since

$$\mathrm{Tr}(\Gamma^d) \le \sum_{j\ge1} \gamma_j < \infty, \qquad \mathrm{Tr}((\Gamma^d)^2) \le \sum_{j\ge1} \gamma_j^2 < \infty,$$

we obtain

$$\mathbb{E}\|\widehat{X}_t^d - \widehat{X}_s^d\|^4 \le C_{\mathrm{inc},4}\left(1 + T^2 + \frac{1}{T^2}\right)(t-s)^2,$$

with $C_{\mathrm{inc},4}$ being a constant independent of d and T. □

Spatial Freezing Defect. We now estimate

$$(\Gamma^d)^{1/2}\big(G_t^d(\widehat{X}_t^d) - G_t^d(\widehat{X}_s^d)\big).$$

Since

$$G_t^d(x) = \sum_{i\in I} p_{i,t}^d(x) c_{i,t}^d(x),$$

its spatial variation has two contributions: an *affine-correction increment*

$$\sum_{i\in I} p_{i,t}^d(x)\big(c_{i,t}^d(x) - c_{i,t}^d(y)\big)$$

and a *responsibility increment*

$$\sum_{i\in I}\big(p_{i,t}^d(x) - p_{i,t}^d(y)\big)c_{i,t}^d(y).$$

The first contribution is controlled by the affine coefficients A_j. The second requires a Lipschitz estimate for the responsibilities $p_{i,t}^d$. We derive this estimate from pairwise differences of the component scores $s_{i,t}^d - s_{\ell,t}^d$ in the next lemma.

Lemma B.8. *For every i, ℓ, j,*

$$s_{i,t,j}^d(x) - s_{\ell,t,j}^d(x) = -a_{i\ell,j}(t)x_j + b_{i\ell,j}(t),$$

where

$$a_{i\ell,j}(t) := \frac{1}{v_{i,t,j}} - \frac{1}{v_{\ell,t,j}}, \qquad b_{i\ell,j}(t) := \frac{m_{ij}}{v_{i,t,j}} - \frac{m_{\ell j}}{v_{\ell,t,j}}.$$

Moreover,

$$|a_{i\ell,j}(t)| \le A_j, \qquad |b_{i\ell,j}(t)| \le B_j.$$

Hence, with

$$\mathcal{A}^d(x) := \left(\sum_{j=1}^d A_j^2 x_j^2\right)^{1/2}, \qquad B_0 := \left(2\sum_{j\ge 1} B_j^2\right)^{1/2},$$

we have, uniformly in i, ℓ, d, t, x,

$$\|s_{i,t}^d(x) - s_{\ell,t}^d(x)\| \le \sqrt{2}\,\mathcal{A}^d(x) + B_0.$$

Proof. The affine formula is immediate from the definition of the component scores $s_{i,t}^d$. Also, since $v_{i,t,j}, v_{\ell,t,j} \ge \underline{\sigma}_j$,

$$|a_{i\ell,j}(t)| = \left|\frac{1}{v_{i,t,j}} - \frac{1}{v_{\ell,t,j}}\right| = \frac{|\sigma_{\ell j} - \sigma_{ij}|}{v_{i,t,j}v_{\ell,t,j}} \le \frac{\delta\sigma_j}{\underline{\sigma}_j^2} = A_j.$$

Similarly,

$$|b_{i\ell,j}(t)| \le \frac{|m_{ij} - m_{\ell j}|}{v_{i,t,j}} + |m_{\ell j}|\left|\frac{1}{v_{i,t,j}} - \frac{1}{v_{\ell,t,j}}\right| \le \frac{\Delta m_j}{\underline{\sigma}_j} + \frac{\overline{m}_j\delta\sigma_j}{\underline{\sigma}_j^2} = B_j.$$

The norm bound follows by expanding the square, using

$$(a-b)^2 \le 2a^2 + 2b^2,$$

and applying the bounds on $a_{i\ell,j}(t)$ and $b_{i\ell,j}(t)$:

$$\|s_{i,t}^d(x) - s_{\ell,t}^d(x)\|^2 \le 2\sum_{j=1}^d A_j^2x_j^2 + 2\sum_{j=1}^d B_j^2 \le 2(\mathcal{A}^d(x))^2 + B_0^2.$$

□

The previous lemma is the missing ingredient for the responsibility part of the spatial variation. Together with the affine bounds in Lemma B.4, it allows us to estimate

$$G_t^d(\widehat{X}_t^d) - G_t^d(\widehat{X}_s^d).$$

The increment bound from Lemma B.7 then converts this pointwise spatial estimate into a bound of order $\left(1 + T + T^{-1}\right)(t-s)$.

Lemma B.9. *Under* (9)*,* (11)*, and* (12)*, there exists a constant C_{sp}, independent of d and T, such that for all $0 \le s \le t \le T$,*

$$\mathbb{E}\left\|(\Gamma^d)^{1/2}\big(G_t^d(\widehat{X}_t^d) - G_t^d(\widehat{X}_s^d)\big)\right\|^2 \le C_{\mathrm{sp}}\left(1 + T + \frac{1}{T}\right)(t-s).$$

Proof. Set

$$C_{A,\gamma} := \sum_{j\ge 1}\gamma_j A_j^2 < \infty, \qquad C_{c,0} := \left(\sum_{j\ge 1}\gamma_j \frac{\overline{m}_j^2}{\underline{\sigma}_j^2}\right)^{1/2}, \qquad C_{c,1} := (2C_{A,\gamma})^{1/2}.$$

Fix d, t and $x, y \in \mathbb{R}^d$. Decompose

$$G_t^d(x) - G_t^d(y) = \sum_{i\in I} p_{i,t}^d(x)\big(c_{i,t}^d(x) - c_{i,t}^d(y)\big) + \sum_{i\in I}\big(p_{i,t}^d(x) - p_{i,t}^d(y)\big)c_{i,t}^d(y).$$

For the affine-correction increment, convexity and Lemma B.4 give

$$\begin{aligned}\left\|(\Gamma^d)^{1/2}\sum_{i\in I} p_{i,t}^d(x)\big(c_{i,t}^d(x) - c_{i,t}^d(y)\big)\right\|^2 &\le \sum_{i\in I} p_{i,t}^d(x)\left\|(\Gamma^d)^{1/2}\big(c_{i,t}^d(x) - c_{i,t}^d(y)\big)\right\|^2 \\ &\le \sum_{j=1}^d \gamma_j A_j^2 (x_j - y_j)^2.\end{aligned}$$

Hence, taking $x = \widehat{X}_t^d$ and $y = \widehat{X}_s^d$,

$$\mathbb{E}\sum_{j=1}^d \gamma_j A_j^2 (\widehat{X}_{t,j}^d - \widehat{X}_{s,j}^d)^2 \le C_{A,\gamma}\mathbb{E}\|\widehat{X}_t^d - \widehat{X}_s^d\|^2.$$

By Lemma B.7,

$$\mathbb{E}\|\widehat{X}_t^d - \widehat{X}_s^d\|^2 \le \left(\mathbb{E}\|\widehat{X}_t^d - \widehat{X}_s^d\|^4\right)^{1/2} \le C\left(1 + T + \frac{1}{T}\right)(t-s),$$

which gives the desired spatial bound for the affine-correction increment.

We next control the responsibility increment. By Lemma B.4,

$$\sup_{i\in I}\|(\Gamma^d)^{1/2}c_{i,t}^d(y)\| \le \sqrt{2}\,C_{c,0} + C_{c,1}\|y\|.$$

Moreover,

$$\nabla p_{i,t}^d(z) = p_{i,t}^d(z)\left(s_{i,t}^d(z) - \sum_{\ell\in I} p_{\ell,t}^d(z)s_{\ell,t}^d(z)\right).$$

Therefore, using Lemma B.8 and convexity,

$$\sum_{i\in I}\|\nabla p_{i,t}^d(z)\| \le \sum_{i,\ell\in I} p_{i,t}^d(z)p_{\ell,t}^d(z)\|s_{i,t}^d(z) - s_{\ell,t}^d(z)\| \le \sqrt{2}\,\mathcal{A}^d(z) + B_0.$$

The mean-value theorem along the segment $z_\theta = (1-\theta)y + \theta x$ gives

$$\begin{aligned}\sum_{i\in I}|p_{i,t}^d(x) - p_{i,t}^d(y)| &\le \|x-y\|\int_0^1 \sum_{i\in I}\|\nabla p_{i,t}^d(z_\theta)\|\,d\theta \\ &\le \big(\sqrt{2}\,\mathcal{A}^d(x) + \sqrt{2}\,\mathcal{A}^d(y) + B_0\big)\|x-y\|.\end{aligned}$$

Hence

$$\left\|(\Gamma^d)^{1/2}\sum_{i\in I}\big(p^d_{i,t}(x)-p^d_{i,t}(y)\big)c^d_{i,t}(y)\right\| \\ \le \big(\sqrt{2}\,C_{c,0}+C_{c,1}\|y\|\big)\big(\sqrt{2}\,\mathcal{A}^d(x)+\sqrt{2}\,\mathcal{A}^d(y)+B_0\big)\|x-y\|.$$

Now set $x=\widehat{X}^d_t$ and $y=\widehat{X}^d_s$. Hölder's inequality with exponents $(4,4,2)$ yields

$$\mathbb{E}\left\|(\Gamma^d)^{1/2}\sum_{i\in I}\big(p^d_{i,t}(\widehat{X}^d_t)-p^d_{i,t}(\widehat{X}^d_s)\big)c^d_{i,t}(\widehat{X}^d_s)\right\|^2 \\ \le \left\|\sqrt{2}\,C_{c,0}+C_{c,1}\|\widehat{X}^d_s\|\right\|^2_{L^8}\left\|\sqrt{2}\,\mathcal{A}^d(\widehat{X}^d_t)+\sqrt{2}\,\mathcal{A}^d(\widehat{X}^d_s)+B_0\right\|^2_{L^8}\left\|\widehat{X}^d_t-\widehat{X}^d_s\right\|^2_{L^4}.$$

The first factor is uniformly bounded by Lemma B.1, with no dependence on T. To bound the second factor, note that

$$\big(\mathcal{A}^d(x)\big)^2=\sum_{j=1}^d A_j^2x_j^2.$$

By Minkowski's inequality in L^4,

$$\Big(\mathbb{E}\big[\mathcal{A}^d(\widehat{X}^d_r)\big]^8\Big)^{1/4}\le\sum_{j=1}^d A_j^2\left(\mathbb{E}|\widehat{X}^d_{r,j}|^8\right)^{1/4}.$$

Under $\mathrm{Law}(\widehat{X}^d_r)=\rho^d_r$, the j-th coordinate is a one-dimensional Gaussian mixture with component means bounded by $\overline{m}_j$ and variances bounded by $\overline{\sigma}_j+\lambda_j$. Hence there is a constant $C>0$ such that

$$\left(\mathbb{E}|\widehat{X}^d_{r,j}|^8\right)^{1/4}\le C(\overline{\sigma}_j+\lambda_j+\overline{m}_j^2).$$

By (12),

$$\sum_{j\ge1}A_j^2(\overline{\sigma}_j+\lambda_j+\overline{m}_j^2)<\infty.$$

Therefore the second factor is uniformly bounded, again with no dependence on T. Finally, Lemma B.7 gives

$$\left\|\widehat{X}^d_t-\widehat{X}^d_s\right\|^2_{L^4}\le C\left(1+T+\frac{1}{T}\right)(t-s).$$

Combining the affine and responsibility estimates proves the result for some constant C_{sp}. □

Temporal Freezing Defect. We next estimate

$$(\Gamma^d)^{1/2}\big(G^d_t(\widehat{X}^d_s)-G^d_s(\widehat{X}^d_s)\big),$$

where the spatial variable is fixed and only the time argument changes. Using

$$G^d_t(x)=\sum_{i\in I}p^d_{i,t}(x)c^d_{i,t}(x),$$

we split the temporal variation into two contributions:

$$\sum_{i\in I}p^d_{i,t}(x)\big(c^d_{i,t}(x)-c^d_{i,s}(x)\big),$$

the affine-correction increment, and

$$\sum_{i\in I}\big(p^d_{i,t}(x)-p^d_{i,s}(x)\big)c^d_{i,s}(x),$$

the responsibility increment. The first term is controlled by the time derivatives of the affine-correction coefficients, namely $\dot{\alpha}_{ij}$ and $\dot{\beta}_{ij}$. The second term requires a bound on the time variation of the responsibilities, which is obtained through

$$\zeta^d_{i,t}(x):=\partial_t\log\varphi^d_{i,t}(x).$$

Since $\kappa_t=(T-t)/T$, each time derivative contributes a factor $1/T$. This is the mechanism behind the final bound of order $(t-s)^2/T^2$.

Lemma B.10. *Define*

$$\zeta_{i,t}^d(x) := \partial_t \log \varphi_{i,t}^d(x), \qquad \mathcal{D}^d(x) := \sum_{j=1}^d D_j x_j^2.$$

Under (9)*,* (13)*, and* (14)*, there exist constants H_0, H_1, independent of d and T, such that, uniformly in i, ℓ, d, t, x,*

$$|\zeta_{i,t}^d(x) - \zeta_{\ell,t}^d(x)| \le \frac{1}{T}\left(H_0 + H_1\|x\| + \mathcal{D}^d(x)\right).$$

Proof. A direct differentiation gives

$$\zeta_{i,t}^d(x) = \frac{1}{2T}\sum_{j=1}^d \lambda_j \left[\frac{1}{v_{i,t,j}} - \frac{(x_j - m_{ij})^2}{v_{i,t,j}^2}\right].$$

Therefore

$$\zeta_{i,t}^d(x) - \zeta_{\ell,t}^d(x) = \frac{1}{2T}\sum_{j=1}^d \lambda_j \left[\left(\frac{1}{v_{i,t,j}} - \frac{1}{v_{\ell,t,j}}\right) - \left(\frac{(x_j - m_{ij})^2}{v_{i,t,j}^2} - \frac{(x_j - m_{\ell j})^2}{v_{\ell,t,j}^2}\right)\right].$$

We expand the quadratic difference:

$$\frac{(x_j - m_{ij})^2}{v_{i,t,j}^2} - \frac{(x_j - m_{\ell j})^2}{v_{\ell,t,j}^2} = \left(\frac{1}{v_{i,t,j}^2} - \frac{1}{v_{\ell,t,j}^2}\right)x_j^2 - 2\left(\frac{m_{ij}}{v_{i,t,j}^2} - \frac{m_{\ell j}}{v_{\ell,t,j}^2}\right)x_j + \left(\frac{m_{ij}^2}{v_{i,t,j}^2} - \frac{m_{\ell j}^2}{v_{\ell,t,j}^2}\right).$$

The coefficient of x_j^2 is bounded by

$$\left|\frac{1}{v_{i,t,j}^2} - \frac{1}{v_{\ell,t,j}^2}\right| \le \frac{2\delta\sigma_j}{\underline{\sigma}_j^3},$$

and hence, after multiplication by λ_j, by $2D_j$.

For the linear coefficient,

$$\left|\frac{m_{ij}}{v_{i,t,j}^2} - \frac{m_{\ell j}}{v_{\ell,t,j}^2}\right| \le \frac{\Delta m_j}{\underline{\sigma}_j^2} + \frac{2\overline{m}_j\delta\sigma_j}{\underline{\sigma}_j^3} \le \frac{2B_j}{\underline{\sigma}_j}.$$

Thus

$$\sum_{j=1}^d \lambda_j \left|\frac{m_{ij}}{v_{i,t,j}^2} - \frac{m_{\ell j}}{v_{\ell,t,j}^2}\right| |x_j| \le 2\left(\sum_{j\ge 1}\lambda_j^2 \frac{B_j^2}{\underline{\sigma}_j^2}\right)^{1/2}\|x\|.$$

For the constant terms, using

$$|m_{ij}^2 - m_{\ell j}^2| \le 2\overline{m}_j\Delta m_j,$$

we get

$$\left|\frac{m_{ij}^2}{v_{i,t,j}^2} - \frac{m_{\ell j}^2}{v_{\ell,t,j}^2}\right| \le \frac{2\overline{m}_j\Delta m_j}{\underline{\sigma}_j^2} + \frac{2\overline{m}_j^2\delta\sigma_j}{\underline{\sigma}_j^3} \le \frac{2\overline{m}_j B_j}{\underline{\sigma}_j}.$$

Hence

$$\sum_{j\ge 1}\lambda_j \frac{\overline{m}_j B_j}{\underline{\sigma}_j} \le \left(\sum_{j\ge 1}\overline{m}_j^2\right)^{1/2}\left(\sum_{j\ge 1}\lambda_j^2\frac{B_j^2}{\underline{\sigma}_j^2}\right)^{1/2} < \infty.$$

Finally,

$$\sum_{j\ge 1}\lambda_j A_j = \sum_{j\ge 1} D_j\underline{\sigma}_j \le \sum_{j\ge 1} D_j\overline{\sigma}_j < \infty$$

by (13). Collecting these bounds proves the claim, with the only dependence on T given by the explicit prefactor $1/T$. □

We now combine the affine-correction estimate with the preceding time-derivative bound for the responsibilities.

Lemma B.11. *Under* (9)*,* (11)*,* (13)*, and* (14)*, there exists a constant C_{tm}, independent of d and T, such that for all $0 \le s \le t \le T$,*

$$\mathbb{E}\left\|(\Gamma^d)^{1/2}\big(G_t^d(\widehat{X}_s^d) - G_s^d(\widehat{X}_s^d)\big)\right\|^2 \le \frac{C_{\mathrm{tm}}}{T^2}(t-s)^2.$$

Proof. Set

$$C_{A,\gamma} := \sum_{j\ge 1}\gamma_j A_j^2 < \infty, \qquad C_{c,1} := (2C_{A,\gamma})^{1/2}, \qquad C_{c,0} := \left(\sum_{j\ge 1}\gamma_j \frac{\overline{m}_j^2}{\underline{\sigma}_j^2}\right)^{1/2}.$$

Fix d and $0 \le s \le t \le T$. For $x \in \mathbb{R}^d$, we write

$$G_t^d(x) - G_s^d(x) = \sum_{i\in I} p_{i,t}^d(x)\big(c_{i,t}^d(x) - c_{i,s}^d(x)\big) + \sum_{i\in I}\big(p_{i,t}^d(x) - p_{i,s}^d(x)\big)c_{i,s}^d(x).$$

We treat the two terms separately.

We first bound the affine-correction increment. By convexity and Lemma B.4,

$$\begin{aligned}
&\left\|(\Gamma^d)^{1/2}\sum_{i\in I} p_{i,t}^d(x)\big(c_{i,t}^d(x) - c_{i,s}^d(x)\big)\right\|^2 \\
&\quad\le \sup_{i\in I}\left\|(\Gamma^d)^{1/2}\big(c_{i,t}^d(x) - c_{i,s}^d(x)\big)\right\|^2 \\
&\quad\le 2\sup_{i\in I}\sum_{j=1}^d \gamma_j|\alpha_{ij}(t) - \alpha_{ij}(s)|^2 x_j^2 + 2\sup_{i\in I}\sum_{j=1}^d \gamma_j|\beta_{ij}(t) - \beta_{ij}(s)|^2.
\end{aligned}$$

By the mean-value theorem and Lemma B.4,

$$|\alpha_{ij}(t) - \alpha_{ij}(s)| \le \frac{2(t-s)}{T}D_j, \qquad |\beta_{ij}(t) - \beta_{ij}(s)| \le \frac{t-s}{T}\frac{\lambda_j\overline{m}_j}{\underline{\sigma}_j^2}.$$

Hence, for $x = \widehat{X}_s^d$,

$$\begin{aligned}
&\mathbb{E}\left\|(\Gamma^d)^{1/2}\sum_{i\in I} p_{i,t}^d(\widehat{X}_s^d)\big(c_{i,t}^d(\widehat{X}_s^d) - c_{i,s}^d(\widehat{X}_s^d)\big)\right\|^2 \\
&\qquad\le \frac{C(t-s)^2}{T^2}\left[\sum_{j\ge 1}\gamma_j D_j^2(\overline{\sigma}_j + \lambda_j + \overline{m}_j^2) + \sum_{j\ge 1}\gamma_j\lambda_j^2\frac{\overline{m}_j^2}{\underline{\sigma}_j^4}\right],
\end{aligned}$$

for some constant $C > 0$, independent of d, s, t, and T. The bracketed quantity is finite by (13) and (14). Therefore the affine-correction increment is bounded by a constant multiple of $(t-s)^2/T^2$, uniformly in d.

We now bound the responsibility increment. Since

$$\partial_r p_{i,r}^d(x) = p_{i,r}^d(x)\left(\zeta_{i,r}^d(x) - \sum_{\ell\in I} p_{\ell,r}^d(x)\zeta_{\ell,r}^d(x)\right),$$

Lemma B.10 gives

$$\sum_{i\in I}|\partial_r p_{i,r}^d(x)| \le \frac{1}{T}\left(H_0 + H_1\|x\| + \mathcal{D}^d(x)\right).$$

Hence

$$\sum_{i\in I}|p_{i,t}^d(x) - p_{i,s}^d(x)| \le \int_s^t \sum_{i\in I}\left|\partial_r p_{i,r}^d(x)\right| \mathrm{d}r \le \frac{t-s}{T}\left(H_0 + H_1\|x\| + \mathcal{D}^d(x)\right).$$

Moreover, Lemma B.4 gives

$$\sup_{i\in I}\|(\Gamma^d)^{1/2}c^d_{i,s}(x)\| \le \sqrt{2}\,C_{c,0} + C_{c,1}\|x\|.$$

Therefore

$$\begin{aligned}
&\left\|(\Gamma^d)^{1/2}\sum_{i\in I}\big(p^d_{i,t}(x) - p^d_{i,s}(x)\big)c^d_{i,s}(x)\right\| \\
&\quad\le \sup_{i\in I}\|(\Gamma^d)^{1/2}c^d_{i,s}(x)\|\sum_{i\in I}|p^d_{i,t}(x) - p^d_{i,s}(x)| \\
&\quad\le \frac{t-s}{T}\big(\sqrt{2}\,C_{c,0} + C_{c,1}\|x\|\big)\big(H_0 + H_1\|x\| + \mathcal{D}^d(x)\big).
\end{aligned}$$

Now set $x = \widehat{X}^d_s$. By Hölder,

$$\begin{aligned}
&\mathbb{E}\left\|(\Gamma^d)^{1/2}\sum_{i\in I}(p^d_{i,t}(\widehat{X}^d_s) - p^d_{i,s}(\widehat{X}^d_s))c^d_{i,s}(\widehat{X}^d_s)\right\|^2 \\
&\quad\le \frac{(t-s)^2}{T^2}\left\|\sqrt{2}\,C_{c,0} + C_{c,1}\|\widehat{X}^d_s\|\right\|^2_{L^4}\left\|H_0 + H_1\|\widehat{X}^d_s\| + \mathcal{D}^d(\widehat{X}^d_s)\right\|^2_{L^4}.
\end{aligned}$$

By Lemma B.1, the moments of $\|\widehat{X}^d_s\|$ needed here are uniformly bounded, with no dependence on T. It remains only to bound $\mathcal{D}^d(\widehat{X}^d_s)$ in L^4. Since

$$\mathcal{D}^d(x) = \sum_{j=1}^d D_j x_j^2,$$

Minkowski's inequality in L^4 gives

$$\left\|D^d(\widehat{X}^d_s)\right\|_{L^4} = \left(\mathbb{E}|\mathcal{D}^d(\widehat{X}^d_s)|^4\right)^{1/4} \le \sum_{j=1}^d D_j\left(\mathbb{E}|\widehat{X}^d_{s,j}|^8\right)^{1/4}.$$

As in the proof of Lemma B.9,

$$\left(\mathbb{E}|\widehat{X}^d_{s,j}|^8\right)^{1/4} \le C(\overline{\sigma}_j + \lambda_j + \overline{m}_j^2),$$

and the summability condition (13) gives

$$\sum_{j\ge 1} D_j(\overline{\sigma}_j + \lambda_j + \overline{m}_j^2) < \infty.$$

Thus the responsibility increment is also bounded by a constant multiple of $(t-s)^2/T^2$, uniformly in d. Combining the affine and responsibility bounds proves the result. □

Path-Space KL Estimate and Conclusion. We now combine the preceding estimates. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, P)$ carry the unique strong solution $\widehat{X}^d$ to (18), together with the driving d-dimensional Brownian motion W^d. Let $P^d := P \circ (\widehat{X}^d)^{-1}$ be the path law of $\widehat{X}^d$ on $C([0,T];\mathbb{R}^d)$, and let Q^d be the path law of the interpolation process Y^d defined by (8). The terminal marginal of P^d is $\rho^d_\star$, whereas the terminal marginal of Q^d is $\mathrm{Law}(Y^d_T)$. Thus a path-space relative entropy estimate, followed by data processing under the evaluation map at time T, gives the desired terminal KL bound.

Proof of Theorem 4.1. Fix $d \ge 1$. By construction, the marginal at time T of Q^d is $\mathrm{Law}(Y^d_T)$, while the marginal at time T of P^d is $\rho^d_\star$.

The two path laws have the same diffusion coefficient $\sqrt{2\Gamma^d}$ and differ only in their drifts. Since the ELP interpolation uses a frozen value at the previous mesh point, it is useful to write the two drifts as adapted functionals on path space. For a continuous path $\omega \in C([0,T];\mathbb{R}^d)$ and $t \in [t_n, t_{n+1})$, set

$$b^{P,d}_t(\omega) := -\Gamma^d(B^d_t)^{-1}\omega_t + \Gamma^d G^d_t(\omega_t) + u^d_t(\omega_t),$$

and

$$b_t^{Q,d}(\omega) := -\Gamma^d (B_t^d)^{-1}\omega_t + \Gamma^d G_{t_n}^d(\omega_{t_n}).$$

The second drift is not Markovian in the current state alone, but it is progressively measurable: on $[t_n, t_{n+1})$ it depends only on the current value ω_t and on the previously observed mesh value ω_{t_n}. Thus it is admissible in the localized path-space Girsanov argument below. Evaluating these functionals along the reference path $\widehat{X}^d$, we obtain

$$(\Gamma^d)^{-1/2}\big(b_t^{P,d}(\widehat{X}^d) - b_t^{Q,d}(\widehat{X}^d)\big) = (\Gamma^d)^{-1/2} u_t^d(\widehat{X}_t^d) + \Delta_t^d,$$

where

$$\Delta_t^d = (\Gamma^d)^{1/2}\Big(G_t^d(\widehat{X}_t^d) - G_{t_n}^d(\widehat{X}_{t_n}^d)\Big).$$

The square integrability needed for Girsanov follows from Lemma B.2, Lemma B.5, and Lemma B.1. Indeed,

$$\|\Delta_t^d\|^2 \le 2\|(\Gamma^d)^{1/2} G_t^d(\widehat{X}_t^d)\|^2 + 2\|(\Gamma^d)^{1/2} G_{t_n}^d(\widehat{X}_{t_n}^d)\|^2.$$

By the growth bound on G_t^d,

$$\|(\Gamma^d)^{1/2} G_t^d(x)\| \le G_0 + G_1\|x\|,$$

and therefore

$$\sup_{d\ge 1}\sup_{t\in[0,T]} \mathbb{E}\|(\Gamma^d)^{1/2} G_t^d(\widehat{X}_t^d)\|^2 \le 2G_0^2 + 2G_1^2 M_2 < \infty.$$

The same estimate applies to $G_{t_n}^d(\widehat{X}_{t_n}^d)$. Hence, for each fixed T,

$$\mathbb{E}\int_0^T \|\Delta_t^d\|^2 \, \mathrm{d}t < \infty.$$

Also,

$$\mathbb{E}\int_0^T \|(\Gamma^d)^{-1/2} u_t^d(\widehat{X}_t^d)\|^2 \, \mathrm{d}t = 4J_{\mathrm{ann}}^d(T) < \infty.$$

Therefore

$$\mathbb{E}\int_0^T \|(\Gamma^d)^{-1/2}\big(b_t^{P,d}(\widehat{X}^d) - b_t^{Q,d}(\widehat{X}^d)\big)\|^2 \, \mathrm{d}t < \infty.$$

Set

$$\beta_t^d := \frac{1}{\sqrt{2}}(\Gamma^d)^{-1/2}\big(b_t^{P,d}(\widehat{X}^d) - b_t^{Q,d}(\widehat{X}^d)\big).$$

Then

$$\mathbb{E}\int_0^T \|\beta_t^d\|^2 \, \mathrm{d}t < \infty.$$

For $m \ge 1$, define the stopping time

$$\tau_m^d := \inf\left\{t \in [0,T] : \int_0^t \|\beta_s^d\|^2 \, ds \ge m\right\} \wedge T.$$

Then

$$\int_0^{\tau_m^d} \|\beta_s^d\|^2 \, ds \le m \qquad P\text{-a.s.}$$

The stopped stochastic exponential

$$Z_t^{d,m} := \exp\left(-\int_0^{t\wedge\tau_m^d} \beta_s^d \cdot dW_s^d - \frac{1}{2}\int_0^{t\wedge\tau_m^d} \|\beta_s^d\|^2 \, ds\right)$$

is therefore a true P-martingale on $[0,T]$. Define a probability measure $Q^{d,m}$ on $(\Omega, \mathcal{F})$ by

$$\frac{dQ^{d,m}}{dP} := Z_T^{d,m}.$$

By the stopped Girsanov theorem,

$$W_t^{d,m} := W_t^d + \int_0^{t\wedge\tau_m^d} \beta_s^d \, ds$$

is a Brownian motion under $Q^{d,m}$, and for $t \le \tau_m^d$,

$$d\widehat{X}_t^d = b_t^{Q,d}(\widehat{X}^d)\, dt + \sqrt{2\Gamma^d}\, dW_t^{d,m}.$$

That is, up to the stopping time τ_m^d, the process $\widehat{X}^d$ solves the same path-dependent equation as the ELP interpolation, with drift $b_t^{Q,d}$ and diffusion coefficient $\sqrt{2\Gamma^d}$. This equation has a unique strong solution, obtained recursively on the mesh intervals: once the value at t_n is known, the drift on $[t_n, t_{n+1})$ is the sum of a time-dependent linear term in the current state and the frozen vector $\Gamma^d G_{t_n}^d(Y_{t_n}^d)$. The same recursive construction gives uniqueness in law for the stopped equation.

Now let

$$S_m : C([0,T];\mathbb{R}^d) \to C([0,T];\mathbb{R}^d), \qquad S_m(\omega) := (\omega_{t\wedge\tau_m(\omega)})_{t\in[0,T]},$$

where $\tau_m(\omega)$ denotes the canonical version of the stopping time above. Write

$$P_m^d := P^d \circ S_m^{-1}, \qquad Q_m^d := Q^d \circ S_m^{-1}.$$

Therefore the law of $\widehat{X}^d_{\cdot\wedge\tau_m^d}$ under $Q^{d,m}$ coincides with Q_m^d.

By data processing under the measurable map $\widehat{X}^d \mapsto \widehat{X}^d_{\cdot\wedge\tau_m^d}$,

$$\mathrm{KL}(P_m^d \,\|\, Q_m^d) \le \mathrm{KL}(P \,\|\, Q^{d,m}).$$

Moreover,

$$\begin{aligned}
\mathrm{KL}(P \,\|\, Q^{d,m}) &= \mathbb{E}_P \log \frac{dP}{dQ^{d,m}} = -\mathbb{E}_P \log Z_T^{d,m} \\
&= \frac{1}{2}\mathbb{E}_P \int_0^{\tau_m^d} \|\beta_t^d\|^2 \, dt \\
&= \frac{1}{4}\mathbb{E}_P \int_0^{\tau_m^d} \left\| (\Gamma^d)^{-1/2} \big(b_t^{P,d}(\widehat{X}^d) - b_t^{Q,d}(\widehat{X}^d) \big) \right\|^2 dt.
\end{aligned}$$

Using $(a+b)^2 \le 2a^2 + 2b^2$ and the definition of $J_{\mathrm{ann}}^d(T)$, we get

$$\mathrm{KL}(P_m^d \,\|\, Q_m^d) \le 2J_{\mathrm{ann}}^d(T) + \frac{1}{2}\sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \mathbb{E}\left[\mathbf{1}_{\{t\le\tau_m^d\}} \|\Delta_t^d\|^2 \right] dt.$$

Since $\tau_m^d \uparrow T$ almost surely, the stopped observations increase to the full path observation. Hence, by monotonicity of relative entropy under increasing observations, followed by monotone convergence on the right-hand side,

$$\mathrm{KL}(P^d \,\|\, Q^d) \le 2J_{\mathrm{ann}}^d(T) + \frac{1}{2}\sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \mathbb{E}\|\Delta_t^d\|^2 \, dt. \tag{21}$$

Finally, data processing under the evaluation map at time T gives

$$\mathrm{KL}\big(\rho_\star^d \,\|\, \mathrm{Law}(Y_T^d)\big) \le \mathrm{KL}(P^d \,\|\, Q^d).$$

It remains to estimate the two terms on the right-hand side of (21). By Lemma B.2,

$$2J_{\mathrm{ann}}^d(T) \le \frac{1}{8T}\sum_{j=1}^d \frac{\lambda_j^2}{\gamma_j \underline{\sigma}_j} \le \frac{1}{8T}\sum_{j\ge1} \frac{\lambda_j^2}{\gamma_j \underline{\sigma}_j}.$$

For the freezing term, fix n and $t \in [t_n, t_{n+1})$. Using the decomposition of Δ_t^d into its spatial and temporal parts,

$$\Delta_t^d = (\Gamma^d)^{1/2}\left(G_t^d(\widehat{X}_t^d) - G_t^d(\widehat{X}_{t_n}^d) \right) + (\Gamma^d)^{1/2}\left(G_t^d(\widehat{X}_{t_n}^d) - G_{t_n}^d(\widehat{X}_{t_n}^d) \right),$$

together with $(a+b)^2 \le 2a^2 + 2b^2$, Lemma B.9, and Lemma B.11, we obtain

$$\mathbb{E}\|\Delta_t^d\|^2 \le 2C_{\mathrm{sp}} \left(1 + T + \frac{1}{T}\right)(t - t_n) + \frac{2C_{\mathrm{tm}}}{T^2}(t - t_n)^2.$$

Integrating over $[t_n, t_{n+1}]$ gives

$$\int_{t_n}^{t_{n+1}} \mathbb{E}\|\Delta_t^d\|^2 \, dt \le C_{\mathrm{sp}} \left(1 + T + \frac{1}{T}\right) h_n^2 + \frac{2C_{\mathrm{tm}}}{3T^2} h_n^3.$$

Summing over n and using

$$\sum_{n=0}^{N-1} h_n^2 \le T h_{\max}, \qquad \sum_{n=0}^{N-1} h_n^3 \le T^2 h_{\max},$$

we find

$$\sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \mathbb{E}\|\Delta_t^d\|^2 \, dt \le \left[C_{\mathrm{sp}}(1 + T + T^2) + \frac{2C_{\mathrm{tm}}}{3}\right] h_{\max}.$$

Inserting this bound into (21) yields

$$\mathrm{KL}\big(\rho_\star^d \,\|\, \mathrm{Law}(Y_T^d)\big) \le \frac{1}{8T} \sum_{j \ge 1} \frac{\lambda_j^2}{\gamma_j \underline{\sigma}_j} + C_{\mathrm{disc},T} h_{\max},$$

where

$$C_{\mathrm{disc},T} := \frac{1}{2} C_{\mathrm{sp}}(1 + T + T^2) + \frac{1}{3} C_{\mathrm{tm}}.$$

In particular, there exists a constant C_{disc}, depending only on the summability bounds, such that

$$C_{\mathrm{disc},T} \le C_{\mathrm{disc}}(1 + T + T^2).$$

The constants do not depend on d or on the mesh. Taking the supremum over d proves Theorem 4.1. The final $\varepsilon_{\mathrm{ann}} + \varepsilon_{\mathrm{disc}}$ statement follows immediately from the estimate above. Since the discretization constant grows at most like $1 + T + T^2$, it is enough to impose

$$h_{\max} \le \frac{\varepsilon_{\mathrm{disc}}}{C_{\mathrm{disc}}(1 + T + T^2)}$$

after choosing T so that the annealing contribution is at most $\varepsilon_{\mathrm{ann}}$. □

B.2 Proof of Proposition 4.1

Let

$$\rho_\star^d = \frac{1}{2}\mathcal{N}(m_1^d, \Sigma_1^d) + \frac{1}{2}\mathcal{N}(m_2^d, \Sigma_2^d), \qquad \rho_0^d = \rho_\star^d * \mathcal{N}(0, C^d).$$

In the example,

$$\sigma_{1j} = \sigma_j = j^{-6}, \qquad \sigma_{2j} = \sigma_j + \delta_j, \qquad \delta_1 = 0, \quad \delta_j = j^{-12} \quad (j \ge 2).$$

Hence

$$\underline{\sigma}_j = \sigma_j = j^{-6}, \qquad \overline{\sigma}_j = \sigma_j + \delta_j, \qquad \overline{\sigma}_j - \underline{\sigma}_j = \delta_j.$$

Moreover,

$$\overline{m}_1 = a, \qquad \overline{m}_j = 0 \quad (j \ge 2).$$

Define

$$\Delta m_j := \sup_{i,\ell \in I} |m_{ij} - m_{\ell j}|.$$

Hence

$$\Delta m_1 = 2a, \qquad \Delta m_j = 0 \quad (j \ge 2).$$

We first verify the summability assumptions (9)–(15). Since

$$\lambda_j = j^{-6}, \qquad \gamma_j = j^{-4}, \qquad \delta_j = j^{-12} \quad (j \ge 2),$$

we have

$$\sum_{j\geq 1}(\overline{\sigma}_j+\lambda_j)\lesssim\sum_{j\geq 1}j^{-6}<\infty,\qquad\sum_{j\geq 1}\overline{m}_j^2=a^2<\infty.$$

Also,

$$\sum_{j\geq 1}\gamma_j=\sum_{j\geq 1}j^{-4}<\infty,$$

and

$$\sum_{j\geq 1}\gamma_j^2\frac{\overline{\sigma}_j+\lambda_j}{\underline{\sigma}_j^2}\lesssim\sum_{j\geq 1}j^{-8}\frac{j^{-6}}{j^{-12}}=\sum_{j\geq 1}j^{-2}<\infty.$$

The term involving the means is finite because $\overline{m}_j=0$ for all $j\geq 2$, while the first coordinate contributes only a constant:

$$\sum_{j\geq 1}\gamma_j^2\frac{\overline{m}_j^2}{\underline{\sigma}_j^2}=\gamma_1^2\frac{a^2}{\sigma_1^2}<\infty.$$

Next,

$$\sum_{j\geq 1}\gamma_j\frac{(\overline{\sigma}_j-\underline{\sigma}_j)^2}{\underline{\sigma}_j^4}=\sum_{j\geq 2}j^{-4}\frac{j^{-24}}{j^{-24}}<\infty,$$

and again

$$\sum_{j\geq 1}\gamma_j\frac{\overline{m}_j^2}{\underline{\sigma}_j^2}<\infty$$

because only the first coordinate contributes.

We now check the conditions controlling the freezing defect. First,

$$\sum_{j\geq 1}\frac{(\overline{\sigma}_j-\underline{\sigma}_j)^2}{\underline{\sigma}_j^4}(\overline{\sigma}_j+\lambda_j+\overline{m}_j^2)\lesssim\sum_{j\geq 2}\frac{j^{-24}}{j^{-24}}\,j^{-6}=\sum_{j\geq 2}j^{-6}<\infty.$$

Furthermore,

$$\sum_{j\geq 1}\left(\frac{\Delta m_j}{\underline{\sigma}_j}+\frac{\overline{m}_j(\overline{\sigma}_j-\underline{\sigma}_j)}{\underline{\sigma}_j^2}\right)^2<\infty,$$

because the summand vanishes for all $j\geq 2$ and the first coordinate is finite.

For the time-variation terms, since $\delta_j=0$ at $j=1$ and $\overline{m}_j=0$ for $j\geq 2$,

$$\sum_{j\geq 1}\frac{\lambda_j(\overline{\sigma}_j-\underline{\sigma}_j)}{\underline{\sigma}_j^3}(\overline{\sigma}_j+\lambda_j+\overline{m}_j^2)\lesssim\sum_{j\geq 2}\frac{j^{-6}j^{-12}}{j^{-18}}\,j^{-6}=\sum_{j\geq 2}j^{-6}<\infty,$$

and

$$\sum_{j\geq 1}\gamma_j\frac{\lambda_j^2(\overline{\sigma}_j-\underline{\sigma}_j)^2}{\underline{\sigma}_j^6}(\overline{\sigma}_j+\lambda_j+\overline{m}_j^2)\lesssim\sum_{j\geq 2}j^{-4}\frac{j^{-12}j^{-24}}{j^{-36}}\,j^{-6}=\sum_{j\geq 2}j^{-10}<\infty.$$

Similarly,

$$\sum_{j\geq 1}\frac{\lambda_j^2}{\underline{\sigma}_j^2}\left(\frac{\Delta m_j}{\underline{\sigma}_j}+\frac{\overline{m}_j(\overline{\sigma}_j-\underline{\sigma}_j)}{\underline{\sigma}_j^2}\right)^2<\infty,$$

again because only the first coordinate contributes. Also,

$$\sum_{j\geq 1}\gamma_j\lambda_j^2\frac{\overline{m}_j^2}{\underline{\sigma}_j^4}<\infty$$

for the same reason. Finally,

$$\sum_{j\geq 1}\frac{\lambda_j^2}{\gamma_j\underline{\sigma}_j}=\sum_{j\geq 1}\frac{j^{-12}}{j^{-4}j^{-6}}=\sum_{j\geq 1}j^{-2}<\infty.$$

Thus all assumptions (9)–(15) hold.

We next prove that

$$\mathrm{KL}(\rho_\star^d \,\|\, \rho_0^d) \to \infty.$$

We use the variational characterization of relative entropy. For $t > 0$, define

$$S_d(x) := \sum_{j=2}^{d} \frac{x_j^2}{\sigma_j}.$$

Then

$$\mathrm{KL}(\rho_\star^d \,\|\, \rho_0^d) \geq -t\, \mathbb{E}_{\rho_\star^d}[S_d] - \log \mathbb{E}_{\rho_0^d}\left[e^{-tS_d}\right].$$

Under $\rho_\star^d$, the tail coordinates have, conditionally on the mixture component, variances

$$\sigma_j \quad \text{or} \quad \sigma_j + \delta_j.$$

Therefore

$$\mathbb{E}_{\rho_\star^d}[S_d] = \frac{1}{2}\sum_{j=2}^{d} 1 + \frac{1}{2}\sum_{j=2}^{d}\left(1 + \frac{\delta_j}{\sigma_j}\right) = (d-1) + \frac{1}{2}\sum_{j=2}^{d} j^{-6} \leq (d-1) + C_0,$$

where $C_0 := \frac{1}{2}\sum_{j\geq 2} j^{-6} < \infty$.

On the other hand, under ρ_0^d, the two tail covariance profiles are

$$\sigma_j + \lambda_j = 2\sigma_j, \qquad \sigma_j + \delta_j + \lambda_j = 2\sigma_j + \delta_j.$$

Hence

$$\mathbb{E}_{\rho_0^d}\left[e^{-tS_d}\right] = \frac{1}{2}\prod_{j=2}^{d}(1+4t)^{-1/2} + \frac{1}{2}\prod_{j=2}^{d}\left(1 + 2t\left(2 + \frac{\delta_j}{\sigma_j}\right)\right)^{-1/2}.$$

Since $\delta_j/\sigma_j = j^{-6} \geq 0$, the second product is bounded above by the first one. Therefore

$$\mathbb{E}_{\rho_0^d}\left[e^{-tS_d}\right] \leq (1+4t)^{-(d-1)/2}.$$

Hence,

$$\mathrm{KL}(\rho_\star^d \,\|\, \rho_0^d) \geq (d-1)\left(\frac{1}{2}\log(1+4t) - t\right) - tC_0.$$

Choosing, for instance, $t = 1/8$, we get

$$\frac{1}{2}\log(1+4t) - t = \frac{1}{2}\log\left(\frac{3}{2}\right) - \frac{1}{8} > 0.$$

Thus

$$\mathrm{KL}(\rho_\star^d \,\|\, \rho_0^d) \to \infty \qquad \text{as } d \to \infty.$$

Finally, since the summability assumptions (9)–(15) have been verified, Theorem 4.1 applies. In the example,

$$\sum_{j\geq 1} \frac{\lambda_j^2}{\gamma_j \underline{\sigma}_j} = \sum_{j\geq 1} j^{-2}.$$

Therefore, for every $T > 0$ and every mesh,

$$\sup_{d\geq 1} \mathrm{KL}\big(\rho_\star^d \,\|\, \mathrm{Law}(Y_T^d)\big) \leq \frac{1}{8T}\sum_{j\geq 1} j^{-2} + C_{\mathrm{disc}}(1+T^2)\, h_{\max} < \infty.$$

C Preconditioner Design for the ELP Scheme in the Power-Law Regime

We use a simple power-law regime to extract a concrete design rule for the ELP scheme from the sufficient conditions of Theorem 4.1. In the simplified setting below, these conditions reveal a clear trade-off.

The role of the preconditioner (γ_j) is two-sided. On the one hand, larger coefficients γ_j accelerate the annealing dynamics and improve the annealing contribution. On the other hand, if γ_j is too large in the high-frequency tail, discretization errors can accumulate across coordinates. Thus the preconditioner must be large enough to make annealing efficient, but sufficiently damped in the tail to control the discretization error.

The main takeaway of this section is the following. In the common-covariance tail regime considered below, these two roles of the preconditioner are balanced by the choice

$$\gamma_j \asymp \lambda_j^{2/3}.$$

With this choice, the dominant annealing and discretization contributions have the same per-coordinate order,

$$\frac{\lambda_j^{4/3}}{\sigma_j}.$$

Thus, in this simplified regime, the sufficient conditions of Theorem 4.1 reduce to

$$\sum_{j\geq 1} \frac{\lambda_j^{4/3}}{\sigma_j} < \infty.$$

For power-law spectra

$$\sigma_j \asymp j^{-a}, \qquad \lambda_j \asymp j^{-b}, \qquad b \geq a,$$

this becomes

$$b > \frac{3(a+1)}{4}.$$

In particular, when the smoothing covariance and the target covariance have the same tail order, $\lambda_j \asymp \sigma_j \asymp j^{-a}$, the balanced preconditioner is

$$\gamma_j \asymp j^{-2a/3},$$

and the sufficient conditions reduce to $a > 3$.

Verification of the Sufficient Conditions. We now verify these claims directly from the conditions of Theorem 4.1.

Proposition C.1. *Consider a Gaussian mixture with common diagonal covariance,*

$$\Sigma_i^d = \Sigma^d = \mathrm{Diag}(\sigma_1, \ldots, \sigma_d), \qquad i \in I,$$

and assume that the component means are supported on finitely many coordinates: there exists $J < \infty$ such that

$$m_{ij} = 0, \qquad i \in I, \quad j > J.$$

Assume that

$$\sigma_j \asymp j^{-a}, \qquad \lambda_j \asymp j^{-b}, \qquad \gamma_j \asymp j^{-c},$$

with $a > 1$, $b \geq a$, and

$$\max\left\{1, \frac{a+1}{2}\right\} < c < 2b - a - 1. \tag{22}$$

Then the conditions (9)–(15) *hold. Consequently, the ELP scheme satisfies the dimension-uniform KL bound of Theorem 4.1.*

Proof. Since the covariance is common across mixture components, we have

$$\underline{\sigma}_j = \overline{\sigma}_j = \sigma_j, \qquad \overline{\sigma}_j - \underline{\sigma}_j = 0.$$

Moreover, because the component means are supported on finitely many coordinates, all terms involving $\overline{m}_j$ or $\sup_{i,\ell} |m_{ij} - m_{\ell j}|$ vanish in the tail. Therefore these terms are harmless for summability.

We first check (9). Since $a > 1$ and $b \geq a$, we have

$$\sum_{j \geq 1} \sigma_j < \infty, \qquad \sum_{j \geq 1} \lambda_j < \infty.$$

Together with the finite-support assumption on the means, this gives (9).

Next consider (10). The condition $\sum_j \gamma_j < \infty$ holds if $c > 1$. For the second condition in (10), since $b \geq a$, we have $\sigma_j + \lambda_j \asymp \sigma_j$ in the tail. Hence

$$\gamma_j^2 \frac{\sigma_j + \lambda_j}{\sigma_j^2} \asymp \frac{\gamma_j^2}{\sigma_j} \asymp j^{-2c+a}.$$

This is summable if

$$2c - a > 1, \qquad \text{that is,} \qquad c > \frac{a+1}{2}.$$

Thus the lower bound in (22) gives (10).

Condition (11) is automatic in the tail because $\overline{\sigma}_j - \underline{\sigma}_j = 0$ and the means are supported on finitely many coordinates. The same argument applies to the covariance-difference and mean-difference terms in (12), (13), and (14).

It remains to check (15). In this setting it reads

$$\sum_{j \geq 1} \frac{\lambda_j^2}{\gamma_j \sigma_j} < \infty.$$

Using the power laws,

$$\frac{\lambda_j^2}{\gamma_j \sigma_j} \asymp j^{-2b+c+a}.$$

Therefore this series is finite if

$$2b - c - a > 1, \qquad \text{that is} \qquad c < 2b - a - 1.$$

Combining this upper bound with the lower bounds above gives exactly (22). Hence all the assumptions (9)–(15) hold, and the conclusion follows from Theorem 4.1. □

Interpretation of the Admissible Preconditioner. The interval (22) has a simple interpretation:

- The lower bound on c is the damping requirement. If c is too small, then $\gamma_j \asymp j^{-c}$ does not decay fast enough, and the high-frequency discretization estimates diverge.
- The upper bound on c is the annealing requirement. If c is too large, then γ_j is too small in the tail, and the annealing contribution

  $$\sum_{j \geq 1} \frac{\lambda_j^2}{\gamma_j \sigma_j}$$

  diverges.

In other words, the preconditioner must be damped enough for discretization, but not so damped that annealing control is lost.

Balanced Preconditioner. The same trade-off can be seen directly, without assuming a power law for γ_j. In this simplified setting, dimension-uniform control follows from controlling the two tail contributions with per-coordinate terms

$$\frac{\gamma_j^2}{\sigma_j} \qquad \text{and} \qquad \frac{\lambda_j^2}{\gamma_j \sigma_j}.$$

The first term comes from the drift and discretization estimates; it penalizes large high-frequency preconditioning. The second term comes from the annealing contribution; it penalizes excessive damping of the preconditioner. Balancing the two terms coordinatewise gives

$$\frac{\gamma_j^2}{\sigma_j} \asymp \frac{\lambda_j^2}{\gamma_j \sigma_j},$$

and hence

$$\gamma_j \asymp \lambda_j^{2/3}.$$

With this balanced choice,

$$\frac{\gamma_j^2}{\sigma_j} \asymp \frac{\lambda_j^2}{\gamma_j \sigma_j} \asymp \frac{\lambda_j^{4/3}}{\sigma_j},$$

and the sufficient conditions of Theorem 4.1 reduce to

$$\sum_{j \geq 1} \frac{\lambda_j^{4/3}}{\sigma_j} < \infty.$$

Under the power laws, this translates into the following proposition.

Proposition C.2. *In the setting of Proposition C.1, assume that*

$$\sigma_j \asymp j^{-a}, \qquad \lambda_j \asymp j^{-b}, \qquad b \geq a.$$

Then the preconditioner that balances the need to accelerate the annealing dynamics with the need to damp high-frequency discretization errors is

$$\gamma_j \asymp \lambda_j^{2/3} \asymp j^{-2b/3}.$$

With this balanced choice,

$$\frac{\gamma_j^2}{\sigma_j} \asymp \frac{\lambda_j^2}{\gamma_j \sigma_j} \asymp \frac{\lambda_j^{4/3}}{\sigma_j} \asymp j^{-4b/3+a}.$$

Therefore the sufficient conditions of Theorem 4.1 are satisfied if

$$\sum_{j \geq 1} j^{-4b/3+a} < \infty,$$

that is

$$b > \frac{3(a+1)}{4}.$$

In particular, if

$$\lambda_j \asymp \sigma_j \asymp j^{-a},$$

the balanced preconditioner is

$$\gamma_j \asymp j^{-2a/3},$$

and the sufficient conditions of Theorem 4.1 reduce to $a > 3$.

D Additional Details on the Numerical Experiments

This appendix collects implementation details and additional diagnostics for the experiments in Section 5. All figures were generated in Google Colab with 13 GB of RAM, with a runtime of about one minute.

Experimental Setup. In Figure 1, we compare Euler–Maruyama (EM) and the exact-linear-part (ELP) discretization on the same preconditioned ALD dynamics. The target is the two-component Gaussian mixture

$$\rho_\star^d = w_1 \mathcal{N}(m_1^d, \Sigma^d) + w_2 \mathcal{N}(m_2^d, \Sigma^d),$$

with

$$w_1 = 0.75, \qquad w_2 = 0.25, \qquad m_1^d = \mathbf{0}, \qquad m_2^d = 8e_1,$$

and common covariance

$$\Sigma^d = \mathrm{Diag}(\sigma_j)_{j=1}^d, \qquad \sigma_j = j^{-6}.$$

The smoothing covariance and preconditioner are diagonal, with

$$\lambda_j = j^{-6}, \qquad \gamma_j = j^{-4}.$$

The annealing path is

$$\rho_t^d = \rho_\star^d * \mathcal{N}(0, \theta(t) C^d), \qquad \theta(t) = 2S(1 - t/T), \qquad S = 5.$$

Both schemes are initialized exactly from

$$\rho_0^d = \rho_\star^d * \mathcal{N}(0, 2SC^d).$$

We use step size $h = 10^{-3}$, 2500 time steps, and hence $T = 2.5$. The dimensions are

$$d \in \{1, 5, 10, 20, 30, 40, 50, 60\}.$$

Spectral Summability Conditions. Since $\lambda_j = \sigma_j$, the initialization becomes increasingly far from the target:

$$\mathrm{KL}(\rho_\star^d \| \rho_0^d)$$

grows linearly with d. At the same time, the ELP summability conditions hold; in particular,

$$\sum_{j \geq 1} \sigma_j < \infty, \qquad \sum_{j \geq 1} \lambda_j < \infty, \qquad \sum_{j \geq 1} \gamma_j < \infty, \qquad \sum_{j \geq 1} \frac{\lambda_j^2}{\gamma_j \sigma_j} < \infty.$$

The two mixture components have the same covariance, so the covariance-mismatch terms in the ELP assumptions vanish.

For EM, however,

$$\frac{\gamma_j}{\sigma_j} = j^2.$$

Thus the largest linearly stable EM step size up to dimension d is of order

$$h_{\max}^{\mathrm{EM}}(d) \simeq \frac{2}{d^2}.$$

Therefore, for a fixed time step, the EM stability condition is violated once the dimension is large enough; in the experiment this happens around $d \approx 45$.

KL Divergence Estimation. The left panel of Figure 1 reports a fixed-k nearest-neighbor estimate of $\mathrm{KL}(\rho_\star^d \| \mathrm{Law}(Y_T^d))$, following the KL estimators of [39, 49]. We use $k = 20$ in the main text. For each dimension, we generate 800 target samples and 800 terminal samples from the discretized sampler.

As a robustness check, we repeat the estimation with $k \in \{10, 20, 30, 50\}$. Figure 2 shows the same behavior across these choices: EM estimate grows rapidly in the stiff regime, whereas the ELP estimate remains stable.

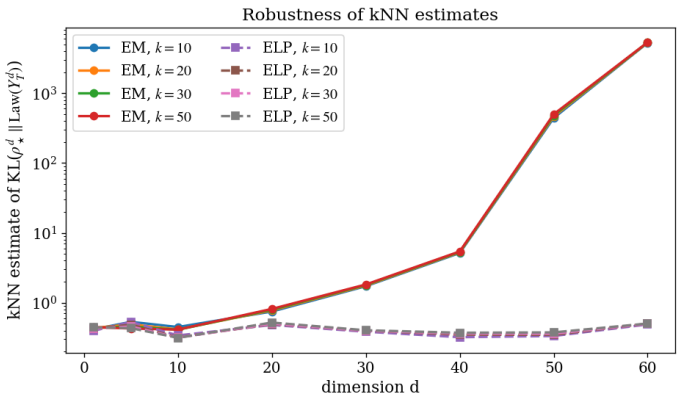

Figure 2: kNN KL estimates for different choices of k. The qualitative behavior is stable across $k \in \{10, 20, 30, 50\}$ (the main text reports $k = 20$): EM grows rapidly once it enters the high-frequency stiffness regime, whereas ELP remains stable.

Variance Profile. The right panel of Figure 1 shows the coordinate-wise variance profile at $d = 50$. For each coordinate j, we compute the empirical variance of the terminal samples and normalize it by the corresponding target marginal variance. This identifies the coordinates in which the instability occurs: EM develops a large variance excess in the high-frequency coordinates, where γ_j/σ_j is largest, whereas ELP remains close to the target scale.

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: Theorem 4.1 derives a dimension-uniform discretization analysis of preconditioned annealed Langevin for Gaussian mixtures.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: We state the assumptions needed for the theoretical results in each proposition, lemma, and theorem. The conclusion also discusses the main limitations.

   Guidelines:

   - The answer [N/A] 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: We state the assumptions needed for the theoretical results in each proposition, lemma, and theorem, as well as throughout the text. The appendix contains the proofs.

Guidelines:

- The answer [N/A] 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: The paper provides the experimental setup in the Appendix.

   Guidelines:

   - The answer [N/A] 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: [No]

Justification: The main contribution of the paper is theoretical; the numerical experiments are intended only to illustrate the theoretical findings. The script to reproduce the results will be made available upon request after acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setting is described in the Appendix.

   Guidelines:

   - The answer [N/A] 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: We report the robustness of the kNN estimation of the KL divergence in the Appendix.

   Guidelines:

   - The answer [N/A] 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: See Appendix D.

   Guidelines:

   - The answer [N/A] 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: We reviewed the NeurIPS Code of Ethics and the research conducted in the paper follows it.

   Guidelines:

   - The answer [N/A] 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: [N/A]

    Justification: The paper develops theoretical results for sampling algorithms. We do not identify direct positive or negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: No external assets were used.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: We do not release new assets.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.