Stochastic Localization via Iterative Posterior Sampling

Louis Grenioux [*1] **Maxence Noble** [*1] **Marylou Gabrié** [1] **Alain Durmus** [1]

Abstract

Building upon score-based learning, new interest in stochastic localization techniques has recently emerged. In these models, one seeks to noise a sample from the data distribution through a stochastic process, called observation process, and progressively learns a denoiser associated to this dynamics. Apart from specific applications, the use of stochastic localization for the problem of sampling from an unnormalized target density has not been explored extensively. This work contributes to fill this gap. We consider a general stochastic localization framework and introduce an explicit class of observation processes, associated with flexible denoising schedules. We provide a complete methodology, *Stochastic Localization via Iterative Posterior Sampling* (SLIPS), to obtain approximate samples of this dynamics, and as a by-product, samples from the target distribution. Our scheme is based on a Markov chain Monte Carlo estimation of the denoiser and comes with detailed practical guidelines. We illustrate the benefits and applicability of SLIPS on several benchmarks of multi-modal distributions, including Gaussian mixtures in increasing dimensions, Bayesian logistic regression and a high-dimensional field system from statistical-mechanics.

1. Introduction

We consider in this paper the problem of sampling from a probability density known up to a normalization constant. This problem finds its origin in various tasks, ranging from Bayesian statistics (Kroese et al., 2011) to statistical mechanics (Krauth, 2006), including now generative modeling (Turner et al., 2019; Grenioux et al., 2023a).

[*]Equal contribution [1]CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France. Correspondence to: Louis Grenioux <louis.grenioux@polytechnique.edu>, Maxence Noble <maxence.noble-bourillot@polytechnique.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Markov chain Monte Carlo (MCMC) samplers are among the most common approaches for this task with a wide span of applicability. In addition, under appropriate conditions on the target, theoretical guarantees can be derived (Dalalyan, 2017; Durmus & Moulines, 2017). However, for complex distributions, simple MCMC algorithms have some limitations. As a workaround, it has been suggested to solve the problem progressively, targeting intermediate smoothed distributions. The resulting algorithms include Replica Exchange (Swendsen & Wang, 1986), Annealed Importance Sampling (Neal, 2001) and Sequential Monte Carlo (Del Moral et al., 2006). Nevertheless, these methods can still largely struggle in high-dimensional settings. To address this issue, approximate inference methods such as Variational Inference (VI) (Wainwright & Jordan, 2008) have emerged, in connection with deep generative models such as Variational Auto-Encoders (Rezende et al., 2014; Kingma & Welling, 2014) or Normalizing Flows (Rezende & Mohamed, 2015).

In contrast to "density-based sampling", which is the setting considered in the present work, generative modeling assumes the availability of training samples and aims to generate similar realizations. Henceforth, we refer to generative modeling as "data-based sampling" as opposed to "density-based sampling". Diffusion-based generative models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) now constitute the state-of-the-art of data-based sampling. In these models, noise is progressively added to the training samples via a forward stochastic process. The derivatives of the logarithm of the forward marginal densities, called *scores*, are learned via score matching techniques (Hyvärinen & Dayan, 2005; Vincent, 2011), and new samples are obtained by simulating the backward process using the learned scores. This approach to generative modeling has been found to scale well with dimension in a large variety of applications (Rombach et al., 2022; Chen et al., 2021; Nichol et al., 2022) and comes with theoretical guarantees (Chen et al., 2023).

Extending diffusion-based model techniques to density-based sampling proves challenging as it requires an efficient estimator of the score without samples of the target. Relying on the link made with stochastic optimal control (Tzen & Raginsky, 2019; De Bortoli et al., 2021; Holdijk et al., 2023; Pavon, 2022), several VI methods using deep

neural networks have recently been proposed for this task (Zhang & Chen, 2022; Berner et al., 2023; Vargas et al., 2023a;c; 2024; Richter et al., 2023). On the other hand, (Huang et al., 2024) take a non-parametric approach and proposed a scheme based on a MCMC estimation of the scores and therefore can potentially mitigate the numerical bias intrinsic in using neural networks or more broadly any parametric estimation method. This is also our approach.

Closely related to denoising diffusion models, Stochastic Localization (SL) techniques have been first employed as a tool to establish results in geometric measure theory (Eldan, 2013; 2020; 2022; Chen & Eldan, 2022), and more recently have been proposed as a density-based sampling approach. (Alaoui et al., 2022; Montanari & Wu, 2023) demonstrate the potential of SL for specific challenging distributions and (Ghio et al., 2023) provide a conjecture on the type of distributions which can be efficiently sampled with these techniques. (Montanari, 2023) provides a blueprint on using SL for density-based sampling but does not discuss a practical strategy for an arbitrary target distribution.

Contributions. Building on these prior works, we bring the following contributions.

- We investigate a general framework of SL that allows to reflect on the role of the signal-to-noise scheduling when using SL for sampling.
- We identify the challenges in its practical implementation and propose a learning-free sampling methodology using MCMC estimation, which requires few tuning. We elucidate our algorithmic design with theoretical and numerical considerations applicable to a certain class of non log-concave distributions.
- We provide numerical evidence of the robustness of the proposed approach in large dimension, beyond the class of distributions amenable to theoretical guarantees. Those results show that the proposed algorithm is on par or superior to modern sampling methods in a wide variety of settings.

The code to reproduce our experiments is available at `https://github.com/h2o64/slips`.

Notation. For any distribution μ with finite first moment, its expectation is denoted by $\mathbf{m}_\mu = \int_{\mathbb{R}^d} x \mathrm{d}\mu(x)$. The density of the Gaussian distribution with mean $\mathbf{m} \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$ is denoted by $x \mapsto \mathsf{N}(x; \mathbf{m}, \Sigma)$.

2. Background on SL for sampling

Consider a target distribution π defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, with $\mathcal{B}(\mathbb{R}^d)$ Borel sets of $\mathbb{R}^d$ endowed with the Euclidean norm. Using SL to sample from π consists in identifying and simulating a stochastic process that converges almost surely to a random variable distributed as π (Alaoui et al., 2022; Montanari & Wu, 2023; Montanari, 2023). A specific example is to consider, given $X \sim \pi$, the stochastic process $(Y_t)_{t\geq 0}$, called the *observation process*, defined by

$$Y_t = tX + \sigma W_t \ , \tag{1}$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion on $\mathbb{R}^d$, independent from X, and $\sigma > 0$[1]. The time-rescaled Y_t/t converges almost surely to X as $t \to \infty$. Moreover, if π admits a finite second order moment, we can show that the 2-Wasserstein distance between the distribution of Y_t/t and π is bounded by $\sigma\sqrt{d/t}$ [2], see Appendix B.2. Thus, for T large, Y_T/T is approximately distributed according to π.

Sampling from $(Y_t)_{t\geq 0}$ using directly (1) cannot be done in practice, precisely since it requires to first sample from π. Nevertheless, this issue can be overcome under the assumption that π admits a finite first moment. Indeed, in this case, $(Y_t)_{t\geq 0}$ solves the Stochastic Differential Equation (SDE),

$$\mathrm{d}Y_t = u_t(Y_t)\mathrm{d}t + \sigma \mathrm{d}B_t \ , \ Y_0 = 0 \ , \tag{2}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion on $\mathbb{R}^d$ and $u_t(y) = \int_{\mathbb{R}^d} x q_t(x|y)\mathrm{d}x$ for any $y \in \mathbb{R}^d$, see (Liptser & Shiryaev, 1977, Theorem 7.12.). The drift function u_t involves the conditional density of X given $Y_t = y \in \mathbb{R}^d$ defined for $x \in \mathbb{R}^d$ as

$$\begin{aligned} q_0(x|y) &\propto \pi(x) \ , \\ q_t(x|y) &\propto \pi(x)\mathsf{N}(x; y/t, \sigma^2/t\,\mathrm{I}_d) \ , \quad t > 0 \ . \end{aligned} \tag{3}$$

In the literature, the *random* conditional expectation $u_t(Y_t) = \mathbb{E}[X|Y_t]$ is often referred to as the *Bayes estimator* of X given Y_t (Robbins, 1992; Saremi & Hyvärinen, 2019) or the *optimal denoiser* of Y_t (Montanari, 2023; Ghio et al., 2023). Moreover, if π has finite second moment, one can show that the SDE (2) admits unique weak solutions (Liptser & Shiryaev, 1977, Theorem 7.6. & Remark 7.2.7.), see Appendix B.2. As a result, one can simulate $(Y_t)_{t\in[0,T]}$ for any $T > 0$ by integrating the SDE (2), while avoiding to first sample from π, if the drift function $(u_t)_{t\in[0,T]}$ (or an approximation of it) is known.

Yet, the challenge of this sampling approach lies in the estimation of $(u_t)_{t\in[0,T]}$ without samples from π available *a priori*. At time $t = 0$, we have $u_0(y) = \mathbf{m}_\pi$. For any $t > 0$, u_t is linked to the *score* of the marginal distribution of Y_t, given by $p_t(y) = \int_{\mathbb{R}^d} \mathsf{N}(y; tx, \sigma^2 t\,\mathrm{I}_d)\mathrm{d}\pi(x)$, following Tweedie's formula given in Appendix A,

$$u_t(y) = y/t + \sigma^2 \nabla \log p_t(y) \ .$$

[1] Note that the authors originally considered $\sigma = 1$.

[2] This bound will be referred to as the *localization rate*.

Therefore, sampling via SL can be reduced to a score estimation task, as noted by (Montanari, 2023), who also establishes a direct connection between the SL induced by the process (1) and variance-preserving diffusions (Song et al., 2021), see Appendix B. In data-based sampling, *i.e.*, when samples $\{X^i\}_{i=1}^N$ from π are at the disposal of the user, u_t can be estimated using a parameterized model through a least-squares regression (Saremi & Hyvärinen, 2019; Montanari, 2023) following score matching techniques.

A few recent works started to discuss the density-based sampling task using SL or diffusions (Montanari & Wu, 2023; Huang et al., 2024; Vargas et al., 2023a). Here, we focus on the SL formalism and propose in Section 4 a sampling algorithm for arbitrary distributions based on a learning-free estimator of the denoiser. Before, we provide in Section 3 a working definition of SL with flexible denoising schedules.

3. SL with flexible denoising scheduling

Inspired by recent developments studying optimal noise scheduling in diffusion models (Nichol & Dhariwal, 2021; Kingma et al., 2021; Karras et al., 2022; Kingma & Gao, 2023), we consider a new class of explicit observation processes associated with flexible denoising schedules.

3.1. A more general observation process

For a possibly finite time horizon $T_{\text{gen}} \in (0, \infty]$, we now consider the observation process $(Y_t^\alpha)_{t\in[0,T_{\text{gen}})}$ defined by

$$Y_t^\alpha = \alpha(t)X + \sigma W_t \ , \tag{4}$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion on $\mathbb{R}^d$, independent from $X \sim \pi$, and $\alpha(t) = t^{1/2}g(t)$ with the following assumptions on g:

(a) $g \in \mathrm{C}^0([0, T_{\text{gen}}), \mathbb{R}_+) \cap \mathrm{C}^1((0, T_{\text{gen}}), \mathbb{R}_+)$;

(b) $g(t) \sim Ct^{\beta/2}$ as $t \to 0$, for some $\beta \geq 1$, $C > 0$;

(c) g is strictly increasing and $\lim_{t\to T_{\text{gen}}} g(t) = \infty$.

Under these assumptions, the observation process (4) verifies $Y_0^\alpha = 0$ and $Y_t^\alpha/\alpha(t) \to X$ almost surely as $t \to T_{\text{gen}}$. Moreover, by denoting π_t^α the distribution of $Y_t^\alpha/\alpha(t)$, one can show that for any $t \in (0, T_{\text{gen}})$, the localization rate is now function of $g(t)$:

$$W_2(\pi, \pi_t^\alpha) \leq \sigma\sqrt{d}/g(t) \ , \tag{5}$$

where W_2 denotes the 2-Wasserstein distance. Proofs and refinements in the case where π is a Gaussian distribution are postponed to Appendix B.2. We recover the process (1) discussed in Section 2 with $g(t) = t^{1/2}$ and $T_{\text{gen}} = \infty$.

As in the previous section, sampling from the observation process via (4) requires to first sample from π. We again bypass this issue by introducing an equivalent SDE. In Appendix B.2, we indeed show that $(Y_t^\alpha)_{t\in[0,T_{\text{gen}})}$ solves

$$\mathrm{d}Y_t^\alpha = \dot\alpha(t)u_t^\alpha(Y_t^\alpha)\mathrm{d}t + \sigma \mathrm{d}B_t \ , \ Y_0^\alpha = 0 \ , \tag{6}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion on $\mathbb{R}^d$ and the drift function u_t^α is defined for any $y \in \mathbb{R}^d$ by

$$u_t^\alpha(y) = \textstyle\int_{\mathbb{R}^d} x q_t^\alpha(x|y)\mathrm{d}x \ . \tag{7}$$

Here, the conditional density of X given $Y_t = y$ is

$$\begin{aligned} &q_0^\alpha(x|y) \propto \pi(x) \ , \\ &q_t^\alpha(x|y) \propto \pi(x)\mathsf{N}(x; y/\alpha(t), \sigma^2/g(t)^2\,\mathrm{I}_d) \ , \ t \in (0, T_{\text{gen}}) \ . \end{aligned} \tag{8}$$

Assuming further that π has finite second moment and that g satisfies technical conditions, (6) admits unique weak solutions, see Appendix B.2. Then again, if u_t^α is known, integrating (6) up to time $T \in (0, T_{\text{gen}})$ is equivalent to sampling from $(Y_t^\alpha)_{t\in[0,T]}$. Then, $Y_T^\alpha/\alpha(T)$ is approximately distributed according to π for any T close enough to T_{gen}.

The challenge still lies in the estimation of the drift function u_t^α which verifies $u_0^\alpha(y) = \mathbf{m}_\pi$ at time $t = 0$, and which can be expressed, for any $t \in (0, T_{\text{gen}})$, using the score of the marginal distribution $p_t^\alpha(y)$ given by $p_t^\alpha(y) = \int_{\mathbb{R}^d} \mathsf{N}(y; \alpha(t)x, \sigma^2 t\,\mathrm{I}_d)\mathrm{d}\pi(x)$. In Appendix B.2, we indeed show that

$$u_t^\alpha(y) = y/\alpha(t) + \{\sigma^2 t/\alpha(t)\}\nabla \log p_t^\alpha(y) \ , \tag{9}$$

recovering the previously mentioned link between score estimation and denoiser estimation.

Moreover, the score of p_t^α can also be written as an expectation over the same posterior distribution q_t^α but with a different observable

$$\nabla \log p_t^\alpha(y) = \tfrac{1}{\alpha(t)} \textstyle\int_{\mathbb{R}^d} \nabla \log \pi(x) q_t^\alpha(x|y)\mathrm{d}x \ . \tag{10}$$

Hence, this suggests to consider an alternative expression of the SDE defined in (6), involving the expectation given in (10) instead of the denoiser function u_t^α [3]. We refer to Appendix B.3 for more details. However, we experimentally observed this approach was less stable than our original method, which is why we decided to consider the SDE (6).

Table 1: Stochastic Localization schemes.

SL scheme	Hyper-parameters	$g(t)$
Geom-$\infty(\alpha_1)$	$\alpha_1 \geq 1$	$t^{\alpha_1/2}$
Geom(α_1, α_2)	$\alpha_1 \geq 1, \alpha_2 > 0$	$t^{\alpha_1/2}(1-t)^{-\alpha_2/2}$

[3] Equation (10) and its consequence have been independently derived in the concurrent works (Bortoli et al., 2024; Akhound-Sadegh et al., 2024). Both have been publicly available almost at the same time than the present paper.

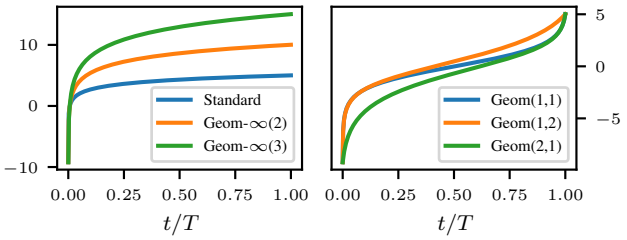

Figure 1: Display of $\log \mathrm{SNR}$ for the localization schemes Geom-∞ (*left*) and Geom (*right*).

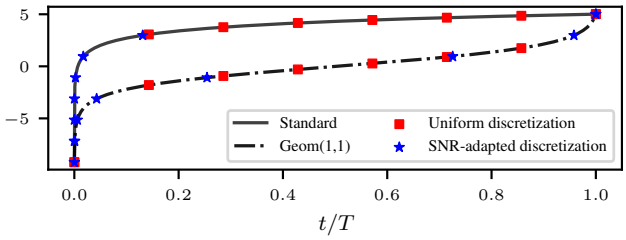

Figure 2: SNR-adapted vs uniform time discretization for the schemes Standard and Geom$(1,1)$. The uniform discretization leads to larger log-SNR differences between timesteps where the SNR increases rapidly.

3.2. Hyper-parameter selection and SNR profile

At first sight, determining appropriate choices for g, σ and T_{gen}, without introducing further complexity, can be challenging. Below, we propose interpretable choices, relying on the analysis of the *signal-to-noise ratio* of the observation process, which determines the speed at which the denoising process localizes on the target distribution.

An intuitive definition of the Signal-to-Noise Ratio (SNR) of the observation process (4) is, for any $t \in [0, T_{\text{gen}})$,

$$\mathrm{SNR}(t) = \frac{\mathbb{E}\left[\|\alpha(t)(X - \mathbb{E}[X])\|^2\right]}{\mathbb{E}\left[\|\sigma W_t\|^2\right]} = \frac{R_\pi^2 g(t)^2}{\sigma^2 d}\,, \quad (11)$$

where $R_\pi^2 = \int_{\mathbb{R}^d} \|x - \mathbf{m}_\pi\|^2 \,\mathrm{d}\pi(x)$ is the scalar variance of π. The monotonicity of g ensures that the SNR is strictly increasing on $[0, T_{\text{gen}})$, with $\lim_{t\to 0} \mathrm{SNR}(t) = 0$ (no information on π at all) and $\lim_{t\to T_{\text{gen}}} \mathrm{SNR}(t) = \infty$ (no noise at all). Further, the scaling in time of the SNR schedule is solely determined by the function g, which also controls the time dependence in the localization rate (5). This observation leads us to fix $\sigma = R_\pi/\sqrt{d}$, assuming that R_π is known, to obtain a SNR profile $\mathrm{SNR}(t) = g^2(t)$ independent of the target distribution or the dimension[4]. In cases where the scalar variance of the target π is unknown, an estimator $\hat{R}_\pi$ can be used instead.

The practical implementation of SL-based sampling requires also the choice of an effective time horizon $T < T_{\text{gen}}$ up to which we integrate the SDE (6), similarly to the maximum denoising time in diffusion models. It needs to be close enough to T_{gen} to consider $Y_T^\alpha/\alpha(T)$ as approximately distributed according to π and as small as possible to ensure computational efficiency. This trade-off is accounted for by selecting T based on reaching a predefined level of the logarithm of the SNR (log-SNR), which is preferable to use in practice. Denoting by η a positive log-SNR threshold, we set $T = T_\eta$, where T_η is such that $\log \mathrm{SNR}(T_\eta) = 2 \log g(T_\eta) = \eta$. In practice, the choice of η does not have a significant impact on performance (see Appendix D.3 Figure 7).

[4]Note that our framework would be unchanged by taking $\sigma = 1$ and defining g up to a multiplicative constant.

Having reduced the set of hyper-parameters to tune, we investigate two main settings for the denoising schedule $g(t)$. Their main characteristics are summarized in Table 1 and examples of corresponding log-SNR profiles are plotted in Figure 1.

(a) **Asymptotic geometric schedule (Geom-∞)**, with $T_{\text{gen}} = \infty$ and $g(t) = t^{\alpha_1/2}$ for $\alpha_1 \geq 1$. This is a natural extension of the observation process discussed in Section 2 for which $\alpha_1 = 1$. We refer to it as the *standard* scheme.

(b) **Non-asymptotic geometric schedule (Geom)**, with $T_{\text{gen}} < \infty$ and $g(t) = (t/T_{\text{gen}})^{\alpha_1/2}(1 - t/T_{\text{gen}})^{-\alpha_2/2}$, for $\alpha_1 \geq 1$ and $\alpha_2 > 0$. In practice, we only consider $T_{\text{gen}} = 1$. If $\alpha_1 = \alpha_2$, the log-SNR profile is similar to the one obtained in diffusion models via cosine noise scheduling (Nichol & Dhariwal, 2021). Moreover, by taking $\alpha_2 < \alpha_1$, the profile becomes flatter near $t = 1$, similarly to the scheduling profiles presented in (Kingma et al., 2021; Kingma & Gao, 2023). Following the framework of (Albergo et al., 2023), $\{(1-t)^{\alpha_2/2} Y_t\}_{t\in[0,1]}$ is a stochastic interpolant between the Dirac mass at 0 and π.

The possibility to adapt the denoising schedule has not yet received attention in the application of SL for sampling. In the following section, we detail the design of our algorithm which adapts to any denoising schedule.

4. SLIPS sampling algorithm

In this section, we consider hyper-parameters σ, T and a denoising schedule g as given in Section 3.2 and expose our strategy for SL-based sampling. As already mentioned, sampling from π via SL consists in integrating the SDE (6) to obtain a realization $\{Y_t^\alpha\}_{t\in[0,T]}$ and set $Y_T^\alpha/\alpha(T)$ as an approximate sample from π. However, the SDE integration faces two main hindrances. First, the drift function $(u_t^\alpha)_{t\in[0,T]}$ defined in (7) is not tractable in general. Second, even if the drift function were known exactly, numerical integration incurs unavoidable discretization errors. We first tackle the latter issue in Section 4.1. We then present a Monte Carlo-based approach to tackle the estimation of u_t^α in Section 4.2. Finally, we present the key idea of our algorithm SLIPS in Section 4.3.

4.1. Handling the discretization error

Consider a time discretization of the interval $[0, T]$ defined by an increasing sequence of timesteps $(t_k)_{k=0}^K$ where $t_0 \geq 0$, $t_K = T$ and $K \geq 1$. We define a sequence of random variables $\{\tilde{Y}_{t_k}^\alpha\}_{k=0}^K$ approximating (6), defined for any $k \in \{0, \dots, K-1\}$ by the recursion

$$\tilde{Y}_{t_{k+1}}^\alpha = \tilde{Y}_{t_k}^\alpha + w_k u_{t_k}^\alpha(\tilde{Y}_{t_k}^\alpha) + \sigma\sqrt{\delta_k} Z_{k+1} \;, \tag{12}$$

where $\tilde{Y}_{t_0}^\alpha = 0$, $\delta_k = t_{k+1} - t_k$, $w_k = \alpha(t_{k+1}) - \alpha(t_k)$ and $(Z_k)_{k=1}^K$ is distributed according to the centered standard Gaussian distribution. Note that (12) results from an *Euler Maruyama* (EM) discretization scheme applied to (6), that corresponds for any $k \in \{0, \dots, K-1\}$ to the exact integration of the SDE

$$\mathrm{d}\tilde{Y}_t^\alpha = \dot{\alpha}(t) u_{t_k}^\alpha(\tilde{Y}_{t_k}^\alpha)\mathrm{d}t + \sigma \mathrm{d}B_t \;,\; t \in [t_k, t_{k+1}] \;.$$

In practice, we rather start the integration from a timestep $t_0 > 0$ for reasons that we detail in the next section. We also adopt a time discretization adapted to the variations of the log-SNR on $[t_0, T]$. Define $\Delta_{\mathrm{SNR}} = \{\log \mathrm{SNR}(T) - \log \mathrm{SNR}(t_0)\}/K$, and for any $k \in \{0, \dots, K\}$, choose t_k such that

$$\log \mathrm{SNR}(t_k) = \log \mathrm{SNR}(t_0) + \Delta_{\mathrm{SNR}} k \;. \tag{13}$$

This discretization defines smaller step-sizes when the variation of $\log \mathrm{SNR}$ is high, *i.e.*, where the denoising process "accelerates" (see Figure 2). We observed that using uniform discretization leads to significantly greater numerical errors in our experiments, see Appendix D.

As described in the next section, we suggest to estimate $u_{t_k}^\alpha(\tilde{Y}_{t_k}^\alpha)$ by Monte Carlo methods. Denoting by $\texttt{MC-Est}(u_t^\alpha(Y))$ an estimator of the drift function at time t evaluated at Y, we come to consider the joint sequence $\{(\tilde{U}_{t_k}^\alpha, \tilde{Y}_{t_k}^\alpha)\}_{k=0}^K$ given by

$$\begin{aligned} \tilde{U}_{t_k}^\alpha &= \texttt{MC-Est}(u_{t_k}^\alpha(\tilde{Y}_{t_k}^\alpha)) \;, \\ \tilde{Y}_{t_{k+1}}^\alpha &= \tilde{Y}_{t_k}^\alpha + w_k \tilde{U}_{t_k}^\alpha + \sigma\sqrt{\delta_k} Z_{k+1} \;, \end{aligned} \tag{14}$$

where $(t_k)_{k=0}^K$ is the SNR-adapted time discretization of $[t_0, T]$ defined in (13). The initialization of $\tilde{Y}_{t_0}^\alpha$ is crucial and will be discussed in Section 4.3. We now detail our Monte Carlo estimation scheme defining $\tilde{U}_{t_k}^\alpha$.

4.2. Estimating the denoiser via MCMC

Monte Carlo estimation with posterior sampling. We aim to build accurate estimators of the quantities

$$u_{t_k}^\alpha(\tilde{Y}_{t_k}^\alpha) = \textstyle\int_{\mathbb{R}^d} x \, q_{t_k}^\alpha(x|\tilde{Y}_{t_k}^\alpha)\mathrm{d}x \;.$$

For ease of notation, we will denote the *random* posterior density $q_{t_k}(\cdot|\tilde{Y}_{t_k}^\alpha)$ by μ_k in this section and recall that for any $k \in \{0, \dots, K\}$

$$\mu_k(x) \propto \pi(x)\mathsf{N}(x; \tilde{Y}_{t_k}^\alpha/\alpha(t_k), \sigma^2/g(t_k)^2 \, \mathrm{I}_d) \;. \tag{15}$$

One possible approach is to estimate $\tilde{U}_{t_k}^\alpha$ with Importance Sampling (IS), by choosing as instrumental distribution $\mathsf{N}(\tilde{Y}_{t_k}^\alpha/\alpha(t_k), \sigma^2/g(t_k)^2 \, \mathrm{I}_d)$. However, IS is known to suffer from large variance in high-dimension unless the proposal is closely adapted to π (Ceriotti et al., 2012). This variance issue translates into sample-size requirements which may grow exponentially with d, see e.g., (Chatterjee & Diaconis, 2018).

Instead, we propose to approximately sample from μ_k with a MCMC algorithm, in the same fashion as (Huang et al., 2024), and compute $\tilde{U}_{t_k}^\alpha$ as the empirical mean of the obtained samples. More precisely, given a Markov chain $\{X_k^j\}_{j=1}^M$ targeting μ_k, we define $\tilde{U}_{t_k}^\alpha = (1/M)\sum_{j=1}^M X_k^j$.

In practice, we turn to the largely used *Metropolis-Adjusted Langevin Algorithm* (MALA) (Roberts & Tweedie, 1996) leveraging gradient-information of the target log-density. Nonetheless, to justify the use of MCMC here, sampling from μ_k has to be easier than the original problem of sampling from π. In particular, we establish in the next paragraph, conditions ensuring that μ_k is log-concave, implying a reasonable mixing time for MALA (Dwivedi et al., 2019).

Main issue in posterior sampling. To illustrate the main challenge in the sampling of μ_k, we focus on a particular setting. In the same spirit as (Saremi et al., 2024, Theorem 1), we consider the following assumption on π.

A0 (Log-concavity outside a compact)**.** *There exist $R > 0$ and $\tau > 0$ such that π is the convolution of μ and $\mathsf{N}(0, \tau^2 \, \mathrm{I}_d)$, where μ is a distribution compactly supported on $\mathsf{B} = \mathrm{B}(\mathbf{m}_\pi, R\sqrt{d})$, i.e., $\mu(\mathbb{R}^d \setminus \mathsf{B}) = 0$.*

This assumption can be equivalently formulated as follows: for any random vector $X \sim \pi$, we have $X = U + G$, where $\|U - \mathbf{m}_\pi\|^2 \leq dR^2$ holds almost surely and $G \sim \mathsf{N}(0, \tau^2 \, \mathrm{I}_d)$ is independent of U. In this case, $d(R^2 + \tau^2)$ is an upper bound on R_π^2, defined as the scalar variance of π in Section 3.2. In particular, this setting includes non log-concave distributions such as Gaussian mixtures, see Appendix C for more details. Under Assumption **A**0, we obtain the following result whose formal statement is given in Appendix E.

Theorem 1. *Assume **A**0. There exists $t_{\mathrm{q}} > 0$ (explicit in Appendix E), depending on g, d, R and τ such that if $t_k > t_{\mathrm{q}}$, μ_k is strongly log-concave. In addition, μ_k is more log-concave as k increases.*

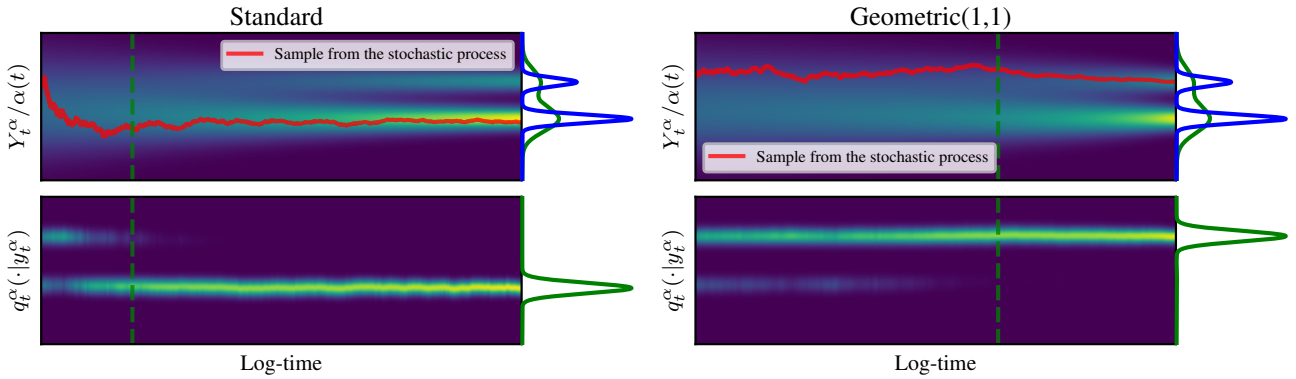

Figure 3: **Duality of log-concavity**: distribution of $Y_t^\alpha/\alpha(t)$ (*up*) and $q_t^\alpha(\cdot|y_t^\alpha)$ (*down*) for $t \in (0, T_{\text{gen}})$, where y_t^α is a realisation of the observation process (red line), for the standard scheme (*left*) and the Geom$(1,1)$ scheme (*right*). The target distribution π is a mixture of two 1D-Gaussian distributions $\mathsf{N}(-2/3, (0.05)^2)$ and $\mathsf{N}(4/3, (0.05)^2)$ with respective weights $2/3$ and $1/3$, which density is given by the blue line. The heat map represents the likelihood of the distributions and the green line on the right edge stands for the distributions taken at the time given by the dotted green line.

This result can be explained by the fact that, for large t_k, the Gaussian term dominates the target term in the posterior (15), since $\sigma^2/g(t_k)^2 \approx 0$ - which intuitively ensures increasing log-concavity. On the other hand, if t_k is close to 0, then μ_k is expected to be close to π, recalling that $q_0^\alpha = \pi$. In this case, for non log-concave target distributions satisfying **A**0, μ_k is non log-concave for small t_k. We illustrate this behaviour in Figure 3 for a bimodal Gaussian mixture: as depicted by the bottom row, the posterior distribution starts multi-modal (non log-concave) and eventually becomes unimodal (log-concave).

Therefore, for a wide variety of target distributions, our MALA-based posterior sampling approach on the time interval $[t_0, T]$ will fail at the very first steps if t_0 is chosen too small. We now explain how to bypass this issue in `SLIPS`.

4.3. Duality of log-concavity is all you need

Initialisation of the integration. Following the previous discussion, a reliable computation of our MCMC-based estimator given $\tilde{U}_{t_k}^\alpha$ requires to have $t_k > t_{\text{q}}$ in order to ensure log-concavity of the random posterior μ_k.

Thus, we aim to start the integration from a $t_0 > t_{\text{q}}$. To initialize the recursion (14) from t_0, one needs to sample the first iterate $\tilde{Y}_{t_0}^\alpha$ distributed according to $p_{t_0}^\alpha$. Given that $\tilde{U}_{t_0}^\alpha$ is a reliable estimator of the denoiser since $t_0 > t_{\text{q}}$, we can derive an approximation of the score $\nabla \log p_{t_0}^\alpha$, via Tweedie's formula (9). We use this estimate to sample from $p_{t_0}^\alpha$ via the *Unadjusted Langevin Algorithm* (ULA) [5] (Roberts & Tweedie, 1996). In this initialization procedure, the successive evaluations of the score $\nabla \log p_{t_0}^\alpha$ in ULA can be computed by applying an inner loop of MALA on $q_{t_0}^\alpha$. This results in a Langevin-within-Langevin scheme, which is presented in Algorithm 2, and detailed in Appendix F.

[5] We do not use MALA since only the score of $p_{t_0}^\alpha$ is available.

For efficient sampling, ULA however requires a condition of log-concavity on $p_{t_0}^\alpha$ (Dalalyan, 2017; Durmus & Moulines, 2017). In Theorem 2, we prove that this property is actually ensured for small values of t_0 under Assumption **A**0.

Theorem 2. *Assume **A**0. There exists $t_{\text{p}} > 0$ (explicit in Appendix E), depending on g, d, R and τ such that if $t < t_{\text{p}}$, p_t^α is strongly log-concave.*

We provide a formal version of this result in Appendix E. Intuitively, when t is small, the Gaussian noise in the observation process (4) overwhelms $\alpha(t)X$, which makes p_t^α log-concave. Note however that this log-concavity property is not verified for large t as p_t^α becomes as log-concave as π by the localization property, see (5). We illustrate this behaviour in the upper row of Figure 3.

Duality of log-concavity. Ideally, we would like to have (i) $t_0 > t_{\text{q}}$ (to ensure that the estimation from $\tilde{U}_{t_k}^\alpha$ is correct) and (ii) $t_0 < t_{\text{p}}$ (to ensure that $p_{t_0}^\alpha$ is log-concave). Under an additional assumption to **A**0, Theorem 3 demonstrates that such t_0 can exist. A proof of this result is given in Appendix E.

Theorem 3. *Assume **A**0, with $dR^2/\tau^2 < 2$. Then, $t_{\text{q}} < t_{\text{p}}$, where t_{q} and t_{p} are given in Theorem 1 and Theorem 2.*

Note however that this extra assumption is restrictive and could be slightly improved, as done for Gaussian mixtures in Appendix E. Then, choosing t_0 boils down to finding a sweet spot where $p_{t_0}^\alpha$ and $q_{t_0}^\alpha$ are both approximately log-concave. Together, these requirements form what we call the *'duality of log-concavity'*. In Figure 3, we show that t_0 can be found for two different localization schemes when considering a Gaussian mixture that does not fit the assumption made in Theorem 3. Furthermore, we highlight in Appendix F.2 that a sweet spot generally exists for a wider variety of target distributions independently of the localization scheme.

In practice, we ensure the log-concavity of $q^\alpha_{t_0}$ first, as the Monte Carlo estimation of $u^\alpha_{t_0}$ is the starting point of the SLIPS algorithm.

4.4. Implementation of SLIPS.

By combining all the previous observations made above, we propose the *Stochastic Localization via Iterative Posterior Sampling* (SLIPS) algorithm, summarized in Algorithm 1, which we detail now.

The SLIPS algorithm has five inputs :

(a) the noising schedule in the observation process $t \mapsto \alpha(t)$ defined on a time interval $[0, T_{\text{gen}})$,

(b) the *initial* integration time in (6): $t_0 \in (0, T_{\text{gen}})$,

(c) the *final* integration time in (6): $T \in (t_0, T_{\text{gen}})$, or equivalently $\eta > 0$ by log-SNR correspondence,

(d) the number of discretization steps in (6): $K \geq 1$,

(e) the number of samples for posterior sampling $M \geq 1$.

The algorithm can be decomposed in two main parts : (i) its initialization (summarized in Algorithm 2) and (ii) the run of the discretized SDE (14) via posterior sampling.

(i) The initialization of SLIPS consists in approximately sampling from $p^\alpha_{t_0}$. Since its score can be expressed via the posterior distribution $q^\alpha_{t_0}$, see (7) and (9), and $q^\alpha_{t_0}$ can itself be sampled from with MCMC methods (see Section 4.2), we propose to use a Langevin-within-Langevin procedure : each step of the outer loop corresponds to an iteration of the Langevin algorithm to sample from $p^\alpha_{t_0}$, and requires local MALA steps to estimate the score by sampling from the posterior $q^\alpha_{t_0}$. The exact computations are detailed in Appendix F. The final iterate of this algorithm is used as the initialization of the second part of SLIPS.

(ii) Once the initialization of SLIPS is done, we turn to the actual core part of SLIPS, which consists in running the discretized SDE given in (14). At step $k+1$ of this recursion (corresponding to the time t_{k+1} in the initial SDE), the denoiser term is approximated by running the MALA algorithm on the posterior distribution $q^\alpha_{t_k}$ conditioned on the k-th iterate of the recursion. Finally, the last iterate of this recursion is considered as an approximate sample from the target distribution.

Complexity of SLIPS. Note that running this procedure is meant to produce one sample from the target distribution. In practice, one may want to produce several samples. In this case, SLIPS can be easily parallelised by running simultaneously independent realizations of the procedure described above. Moreover, we emphasize that the initialization of $\tilde{Y}^\alpha_{t_0}$ is the most challenging step in SLIPS as posterior sampling only gets easier afterwards, by Theorem 1. The computational cost of this initialization, involving a Langevin-within-Langevin procedure, remains however reasonable thanks to a persistent initialization of the ULA and MALA chains (see Appendix F). As such, only a few steps of each is needed in practice (see Appendix H.4).

Limitation of SLIPS. We highlight that t_0 is the only hyper-parameter in our algorithm that requires careful tuning. However, Figure 10 and Figure 11 in Appendix F.2 highlight moderate sensitivity to this hyper-parameter.

Algorithm 1 SLIPS

Input: α, t_0, η, K, M
Set $T = T_\eta$ and $\sigma = \hat{R}_\pi/\sqrt{d}$, see Section 3.2
Set $(t_k)_{k=0}^K$ as the SNR-adapted disc. of $[t_0, T]$, see (13)
Initialize $\tilde{Y}^\alpha_{t_0}$ with Algorithm 2
for $k = 0$ **to** $K-1$ **do**
  Define $\delta_k = t_{k+1} - t_k$, $w_k = \alpha(t_{k+1}) - \alpha(t_k)$
  Simulate $\{X^k_j\}_{j=1}^M \sim \mu_k$ with MALA, see (15)
  Estimate the denoiser by $\tilde{U}^\alpha_{t_k} = (1/M)\sum_{j=1}^M X^k_j$
  Simulate $\tilde{Y}^\alpha_{t_{k+1}} \sim \mathsf{N}(\tilde{Y}^\alpha_{t_k} + w_k \tilde{U}^\alpha_{t_k}, \sigma^2 \delta_k \, \mathrm{I}_d)$
end for
Output: $\tilde{Y}^\alpha_{t_K}/\alpha(t_K)$

Algorithm 2 Langevin-within-Langevin initialization

Input: $\alpha(t_0)$, t_0, σ, N, M
Set $Y^{(0)} \sim \mathsf{N}(0, \sigma^2 t_0 \, \mathrm{I}_d)$ and $\lambda = \sigma^2 t_0/2$
for $n = 0$ **to** $N-1$ **do**
  Simulate $\{X^{(n)}_j\}_{j=1}^M \sim q^\alpha_{t_0}(\cdot|Y^{(n)})$ with MALA
  Estimate the denoiser by $U^{(n)} = (1/M)\sum_{j=1}^M X^{(n)}_j$
  Set $\hat{s}^\alpha_{t_0}(Y^{(n)}) = \left(\alpha(t_0)U^{(n)} - Y^{(n)}\right)/(\sigma^2 t_0)$, see (6)
  Simulate $Y^{(n+1)} \sim \mathsf{N}(Y^{(n)} + \lambda \hat{s}^\alpha_{t_0}(Y^{(n)}), 2\lambda \, \mathrm{I}_d)$
end for
Output: $Y^{(N)}$

5. Related work

Score-based sampling with VI. Building upon score-based generative models and VI, recent works have proposed deep-learning approaches for density-based sampling. Those sampling schemes amount to discretized versions of denoising processes with a parameterized drift function. Two main settings may be distinguished. On one hand, the VI framework is seen as a stochastic optimal control problem involving the time-reversal of the denoising process, see e.g., (Zhang & Chen, 2022; Berner et al., 2023; Vargas et al., 2023b;a). We refer to (Richter et al., 2023) for a global overview of this approach and its extensions. On the other hand, another line of work has proposed parameterized extensions of Annealed Importance Sampling (AIS) (Doucet et al., 2022; Geffner & Domke, 2023; Vargas et al., 2024). Since SLIPS is learning-free, these algorithms are not included in our numerical tests as it would be difficult to draw comparisons at equal computational budget.

Score-based sampling with Monte Carlo. Preliminary study of Monte Carlo score estimation was done by (Huang et al., 2021), followed by (Vargas et al., 2023b), who considered the Föllmer diffusion (Föllmer, 2005; 2006) bridging δ_0 to π in a finite-time setting. However, their method relying on IS does not scale well with dimension, and suffers from numerical unstability. Note that it can be recasted as a particular example of our scheme Geom(1,1), via the stochastic interpolant analogy, see Section 3.2.

Closely related to this work, (Huang et al., 2024) recently proposed *Reverse Diffusion Monte Carlo* (RDMC), a sampling algorithm based on a Monte Carlo estimation of the drift of the time-reversal of a variance preserving diffusion. Many algorithmic choices in RDMC are similar to SLIPS, turning to Langevin for the drift estimation and using Langevin-within-Langevin for initialization. These authors also discuss the choice of the initial integration time, which plays a role similar to our t_0. Notably, under Lipschitz assumption on the score, they derive an upper bound on the overall complexity of RDMC, depending on this time. The present work complements the approach of (Huang et al., 2024). We formalize the crucial trade-off in choosing t_0 according to our notion of 'duality of log-concavity' (see Section 4.3), which is only briefly mentioned in (Huang et al., 2024). This focal point naturally arises from our numerical tests in high-dimension that allows us to assess the potential of SL-based sampling in practice, where (Huang et al., 2024) remained mostly theoretical. A precise comparison is provided in Appendix G.

Multi-Noise Measurements sampling. Recently, (Saremi & Srivastava, 2022) exploited a non-Markovian stochastic process to propose a novel data-based sampling scheme. The Multi-Noise Measurements (MNM) process $(Y^m)_{m=1}^M$ is defined by $Y^m = X + \sigma Z^m$ where $\sigma > 0$, $X \sim \pi$ and Z^m are independently sampled from $\mathsf{N}(0, \mathrm{I}_d)$. Here, the *measurement* Y^m plays the same role as the *observation* in standard SL[6]. In (Saremi et al., 2024), the authors suggest to use this process in the density-based setting by computing $Y^{1:M}$ sequentially through the sampling of Y^m conditioned on $Y^{1:m-1}$. They show that those conditional distributions are increasingly log-concave, see (Saremi et al., 2024, Theorem 1), making their *Once-At-a-Time* (OAT) algorithm efficient. In contrast to our framework, (Saremi et al., 2024) mainly assume that the score of the measurement process is analytically available. Although a Monte Carlo-based approach is proposed to estimate the score in realistic settings, results significantly degrade compared to the analytical case. As we detail in Appendix G, this can be explained by the fact that OAT faces but does not tackle the challenge of 'duality of log-concavity'.

[6]Indeed, both denoising processes have the same localization rate when $M = T$, see Appendix G for more details.

6. Numerical experiments

In this section, we compare SLIPS against SMC, AIS, RDMC (Huang et al., 2024) and OAT (Saremi et al., 2024). Note that we deliberately omit an exhaustive comparison with standard MCMC methods, as they notoriously fail to sample from multi-modal distributions, see Appendix H.1. For SLIPS, we consider three different SL schemes : Standard, Geom(1,1) and Geom(2,1). Except for RDMC, all the algorithms are informed by the scalar variance R_π^2 of the target distribution (or an estimation). We tuned the hyperparameters of each algorithm with coarse grid searches of similar size and similar computational budgets assuming access to an oracle distance metric to the target distribution (see details in Appendix H). For OAT, we estimated the intermediate scores with IS.

Toy target distributions. We first discuss standard target distributions including the *8-Gaussians* ($d = 2$), the *Rings* ($d = 2$) and the *Funnel* ($d = 10$) distributions. Details about their respective definitions are provided in Appendix H. Apart from Funnel, for which we provide results based on the sliced Kolmogorov-Smirnov (KS) distance as per (Grenioux et al., 2023b), we compare the samples obtained with the algorithms against the ground truth using the entropic regularized 2-Wasserstein distance. The first four columns of Table 2 show that SLIPS is on par with most of its competitors on those toy distributions.

Bayesian Logistic Regression. Beyond toy distributions, we sample from the posterior of a Bayesian logistic regression model on two popular datasets : *Ionosphere* ($d = 34$) and *Sonar* ($d = 61$). More details on the design of the model are available in Appendix H. We evaluate the quality of sampling by computing the average predictive log-likelihood of the obtained samples. The last two columns of Table 2 show that, in these higher dimensions, SLIPS is slightly superior to its counterparts, especially OAT and RDMC.

High-dimensional Gaussian Mixtures. As a challenging task, we seek to estimate the relative weight of a bimodal Gaussian mixture with modes $\mathsf{N}(x; -(2/3)\mathbf{1}_d, \Sigma)$ and $\mathsf{N}(x; (4/3)\mathbf{1}_d, \Sigma)$, where $\Sigma = 0.05\,\mathrm{I}_d$ and $\mathbf{1}_d$ is the d-dimensional vector with all components equal to 1, and respective weights $2/3$ and $1/3$. In our experiments, these weights are computed as $\hat{W}$ and $1 - \hat{W}$, where $\hat{W}$ is a Monte Carlo estimation of $\int_{\mathbb{R}^d} \mathbb{1}_{(-\infty,0)^d}(x)\mathrm{d}\pi(x)$. Here, we consider increasing values of d. Figure 4 (*top*) shows that SLIPS recovers the relative weight of the target at a 1% accuracy, even in high dimensions, whereas the other methods fail to give accurate estimates in most dimensions. In Figure 4 (*bottom*), we also illustrate the superiority of SLIPS to capture the local properties of the distribution by displaying the sliced Wasserstein distance to the target distribution. Due to high estimation error, AIS and SMC results are omitted in Figure 4 and displayed in Appendix H.

Table 2: Metrics when sampling toy distributions and posteriors from Bayesian logistic regression models. Bold font indicates best results. The metric for 8-Gaussians and Rings ($d = 2$) is the entropic regularized 2-Wasserstein distance (with regularization hyper-parameter 0.05) (*the lowest, the best*), the metric for Funnel ($d = 10$) is the sliced Kolmogorov-Smirnov distance (*the lowest, the best*) and the metric for the Bayesian logistic regression on Sonar ($d = 34$) and Ionosphere ($d = 61$) datasets corresponds to the average predictive posterior log-likelihood (*the highest, the best*).

ALGORITHM	8-GAUSSIANS (↓)	RINGS (↓)	FUNNEL (↓)	SONAR (↑)	IONOSPHERE (↑)
AIS	1.10 ± 0.09	0.19 ± 0.02	0.037 ± 0.004	-111.04 ± 0.08	-87.92 ± 0.13
SMC	0.99 ± 0.16	0.28 ± 0.06	0.035 ± 0.005	-111.02 ± 0.13	-87.82 ± 0.16
OAT (SAREMI ET AL., 2024)	0.91 ± 0.09	0.18 ± 0.02	0.105 ± 0.005	-280.92 ± 1.34	-205.49 ± 0.61
RDMC (HUANG ET AL., 2024)	1.01 ± 0.05	0.30 ± 0.01	0.082 ± 0.006	-129.88 ± 0.12	-109.84 ± 0.10
SLIPS STANDARD	0.76 ± 0.05	$\mathbf{0.19 \pm 0.01}$	$\mathbf{0.024 \pm 0.003}$	-109.25 ± 0.07	-86.65 ± 0.04
SLIPS GEOM(1,1)	$\mathbf{0.74 \pm 0.12}$	0.20 ± 0.01	0.032 ± 0.002	$\mathbf{-109.14 \pm 0.09}$	$\mathbf{-86.32 \pm 0.10}$
SLIPS GEOM(2,1)	0.75 ± 0.10	0.22 ± 0.02	0.040 ± 0.007	-110.24 ± 0.05	-86.78 ± 0.08

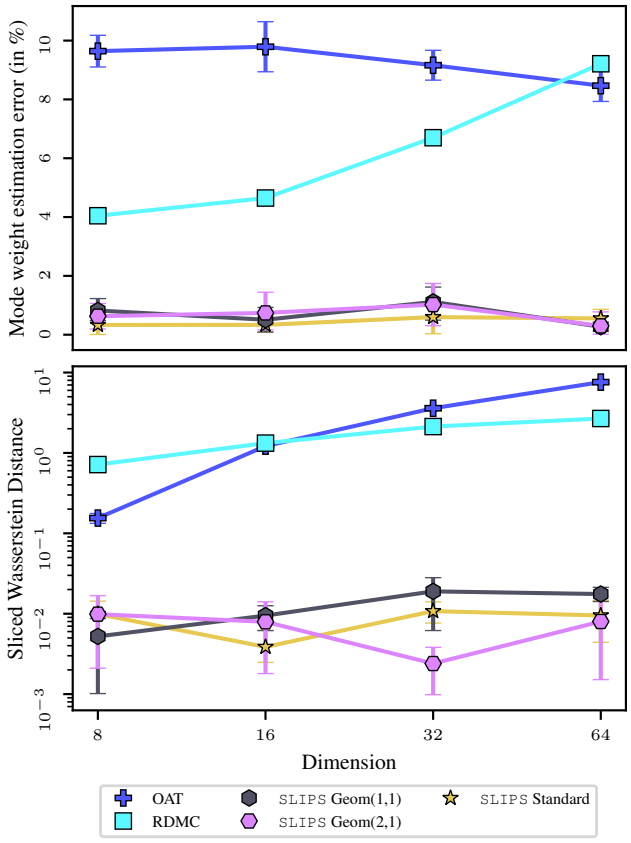

Figure 4: Metrics when sampling a bimodal Gaussian mixture with d growing. **Top**: Relative weight estimation error. **Bottom**: Sliced Wasserstein distance.

Field system ϕ^4 from statistical mechanics. Lastly, we sample the 1D ϕ^4 model, which was recently used as a benchmark in (Gabrié et al., 2022). At the chosen temperature, the distribution has two well distinct modes with relative weight that can be adjusted through a 'local-field' parameter h. We discretize this continuous model with a grid size of 100 (*i.e.*, $d = 100$). We estimate the relative weight between the two modes and compare the results with a Laplace approximation.

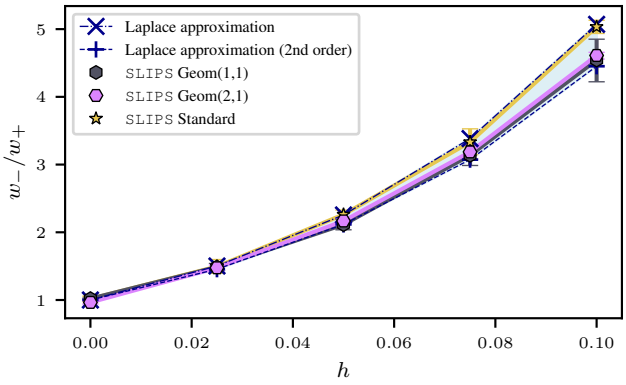

Figure 5: Estimation of the mode weight ratio of ϕ^4 with increasing h - Only SLIPS produced correct samples.

Figure 5 shows that the relative weight estimated by SLIPS lie between the 0-th and 2-nd order Laplace approximations. Due to high estimation error, the results from concurrent algorithms are omitted in Figure 5. We refer to Appendix H for more details on the setting of this numerical experiment.

7. Discussion

In this paper, we introduced a Stochastic Localization (SL) scheme, that features flexible denoising schedule, in order to sample from any unnormalized target density. Relying on this framework, we proposed our algorithm, *Stochastic Localization via Iterative Posterior Sampling* (SLIPS) which leverages Monte Carlo estimation of the SL denoiser. Its design notably reveals the so-called "duality of log-concavity", that can be seen as a trade-off between sampling from the observation process and sampling from the SL posterior. For various localization schemes, we illustrate the performance of SLIPS in high-dimension. In future work, we would like to derive more theoretical guarantees on the phenomenon of duality, investigate improved designs of denoising schedule and consider more challenging target distributions.

Acknowledgements

The authors would like to thank Saeed Saremi and Francis Bach for insightful discussions during the course of this project. The work of LG and MG is supported by Hi! Paris. Part of the work of AD is funded by the European Union (ERC, Ocean, 101071601). Views and opinions expressed here are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Impact Statement

This paper presents work whose goal is to advance the field of sampling from unnormalized densities. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Akhound-Sadegh, T., Rector-Brooks, J., Bose, A. J., Mittal, S., Lemos, P., Liu, C.-H., Sendera, M., Ravanbakhsh, S., Gidel, G., Bengio, Y., Malkin, N., and Tong, A. Iterated denoising energy matching for sampling from boltzmann densities, 2024.

Alaoui, A. E., Montanari, A., and Sellke, M. Sampling from the Sherrington-Kirkpatrick Gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 323–334, 2022. doi: 10.1109/FOCS54457.2022.00038.

Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. Stochastic interpolants: A unifying framework for flows and diffusions, 2023. URL https://arxiv.org/abs/2303.08797.

Berner, J., Richter, L., and Ullrich, K. An optimal control perspective on diffusion-based generative modeling. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=oYIjw37pTP.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Bortoli, V. D., Hutchinson, M., Wirnsberger, P., and Doucet, A. Target score matching, 2024.

Cattiaux, P., Conforti, G., Gentil, I., and Léonard, C. Time reversal of diffusion processes under a finite entropy condition. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 59(4):1844 – 1881, 2023. doi: 10.1214/22-AIHP1320. URL https://doi.org/10.1214/22-AIHP1320.

Ceriotti, M., Brain, G. A. R., Riordan, O., and Manolopoulos, D. E. The inefficiency of re-weighted sampling and the curse of system size in high-order path integration. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 468 (2137):2–17, 2012. doi: 10.1098/rspa.2011.0413. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2011.0413.

Chatterjee, S. and Diaconis, P. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=NsMLjcFaO8O.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Chen, Y. and Eldan, R. Localization schemes: A framework for proving mixing bounds for Markov chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 110–122. IEEE, 2022.

Dalalyan, A. S. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017.

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

Del Moral, P., Doucet, A., and Jasra, A. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006.

Doucet, A., Grathwohl, W., Matthews, A. G., and Strathmann, H. Score-based diffusion meets annealed importance sampling. *Advances in Neural Information Processing Systems*, 35:21482–21494, 2022.

Durmus, A. and Moulines, É. Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis Adjusted Langevin Algorithm. *Statistics and Computing*, 25:5–19, 2015.

Durmus, A. and Moulines, E. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. 2017.

Dwivedi, R., Chen, Y., Wainwright, M. J., and Yu, B. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183): 1–42, 2019.

Eldan, R. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.

Eldan, R. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3-4):737–755, 2020.

Eldan, R. Analysis of high-dimensional distributions using pathwise methods. *Proceedings of ICM, to appear*, 2022.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Föllmer, H. An entropy approach to the time reversal of diffusion processes. In *Stochastic Differential Systems Filtering and Control: Proceedings of the IFIP-WG 7/1 Working Conference Marseille-Luminy, France, March 12–17, 1984*, pp. 156–163. Springer, 2005.

Föllmer, H. Time reversal on Wiener space. In *Stochastic Processes—Mathematics and Physics: Proceedings of the 1st BiBoS-Symposium held in Bielefeld, West Germany, September 10–15, 1984*, pp. 119–129. Springer, 2006.

Gabrié, M., Rotskoff, G. M., and Vanden-Eijnden, E. Adaptive Monte Carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10): e2109420119, 2022.

Gabrié, M., Rotskoff, G. M., and Vanden-Eijnden, E. Adaptive Monte Carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10):e2109420119, 2022. doi: 10.1073/pnas.2109420119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2109420119.

Geffner, T. and Domke, J. Langevin diffusion variational inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 576–593. PMLR, 2023.

Ghio, D., Dandi, Y., Krzakala, F., and Zdeborová, L. Sampling with flows, diffusion and autoregressive neural networks: A spin-glass perspective. In *NeurIPS 2023 Workshop on Diffusion Models*. arXiv, August 2023. URL http://arxiv.org/abs/2308.14085. arXiv:2308.14085 [cond-mat].

Grenioux, L., Moulines, E., and Gabrié, M. Balanced training of energy-based models with adaptive flow sampling. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023a. URL https://openreview.net/forum?id=AwJ2NqxWlk.

Grenioux, L., Oliviero Durmus, A., Moulines, E., and Gabrié, M. On sampling with approximate transport maps. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11698–11733. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/grenioux23a.html.

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

Holdijk, L., Du, Y., Hooft, F., Jaini, P., Ensing, B., and Welling, M. PIPS: Path integral stochastic optimal control for path sampling in molecular dynamics, 2023. URL https://openreview.net/forum?id=TnIZfXSFJAh.

Huang, J., Jiao, Y., Kang, L., Liao, X., Liu, J., and Liu, Y. Schrödinger-Föllmer sampler: sampling without ergodicity. *arXiv preprint arXiv:2106.10880*, 2021.

Huang, X., Dong, H., HAO, Y., Ma, Y., and Zhang, T. Reverse Diffusion Monte Carlo. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kIPEyMSdFV.

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

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577, 2022.

Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Kingma, D. P. and Gao, R. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. *Proc. Int. Conf. on Learning Representations*, 2014.

Krauth, W. Statistical mechanics: algorithms and computations. *OUP Oxford*, 13, 2006.

Kroese, D. P., Taimre, T., and Botev, Z. I. *Handbook of Monte Carlo methods*. Number 706 in Wiley series in probability and statistics. Wiley, Hoboken, 2011. ISBN 978-0-470-17793-8.

Liptser, R. S. and Shiryaev, A. N. *Statistics of random processes: General theory*, volume 394. Springer, 1977.

Montanari, A. Sampling, diffusions, and stochastic localization, 2023.

Montanari, A. and Wu, Y. Posterior sampling from the spiked models via diffusion processes. *arXiv preprint arXiv:2304.11449*, 2023.

Nadjahi, K., Durmus, A., Simsekli, U., and Badeau, R. Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. *Advances in Neural Information Processing Systems*, 32, 2019.

Neal, R. M. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

Neal, R. M. Slice sampling. *The Annals of Statistics*, 31(3): 705 – 767, 2003. doi: 10.1214/aos/1056562461. URL https://doi.org/10.1214/aos/1056562461.

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

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., and Chen, M. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/nichol22a.html.

Olver, F. W. *NIST handbook of mathematical functions*. Cambridge university press, 2010.

Pavon, M. On local entropy, stochastic control, and deep neural networks. *IEEE Control Systems Letters*, 7:437–441, 2022.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *Proceedings of The 31st International Conference in Machine Learning, Beijing China*, 32:1278–, 2014. URL http://jmlr.org/proceedings/papers/v32/rezende14.html%5Cnpapers3://publication/uuid/F2747569-7719-4EAC-A5A7-9ECA9D6A8FE6. arXiv: 1401.4082v3 ISBN: 9781634393973.

Richter, L., Berner, J., and Liu, G.-H. Improved sampling via learned diffusions. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023. URL https://openreview.net/forum?id=uLgYD7ie0O.

Robbins, H. E. An empirical Bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pp. 388–394. Springer, 1992.

Roberts, G. O. and Tweedie, R. L. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. ISSN 13507265. URL http://www.jstor.org/stable/3318418.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Sachs, M., Leimkuhler, B., and Danos, V. Langevin dynamics with variable coefficients and nonconservative forces: from stationary states to numerical methods. *Entropy*, 19 (12):647, 2017.

Saremi, S. and Hyvärinen, A. Neural empirical Bayes. *Journal of Machine Learning Research*, 20(181):1–23, 2019. URL http://jmlr.org/papers/v20/19-216.html.

Saremi, S. and Srivastava, R. K. Multimeasurement generative models. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=QRX0nCX_gk.

Saremi, S., Srivastava, R. K., and Bach, F. Universal smoothed score functions for generative modeling, 2023.

Saremi, S., Park, J. W., and Bach, F. Chain of log-concave Markov chains. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=yiMB2DOjsR.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *The Ninth International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Swendsen, R. H. and Wang, J.-S. Replica Monte Carlo Simulation of Spin-Glasses. *Physical Review Letters*, 57(21):2607–2609, November 1986. doi: 10.1103/PhysRevLett.57.2607. URL https://link.aps.org/doi/10.1103/PhysRevLett.57.2607. Publisher: American Physical Society.

Turner, R., Hung, J., Frank, E., Saatchi, Y., and Yosinski, J. Metropolis-Hastings generative adversarial networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6345–6353. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/turner19a.html.

Tzen, B. and Raginsky, M. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Conference on Learning Theory*, pp. 3084–3114. PMLR, 2019.

Vargas, F., Grathwohl, W. S., and Doucet, A. Denoising diffusion samplers. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=8pvnfTAbu1f.

Vargas, F., Ovsianas, A., Fernandes, D., Girolami, M., Lawrence, N. D., and Nüsken, N. Bayesian learning via neural Schrödinger–Föllmer flows. *Statistics and Computing*, 33(1):3, 2023b.

Vargas, F., Reu, T., and Kerekes, A. Expressiveness remarks for denoising diffusion based sampling. In *Fifth Symposium on Advances in Approximate Bayesian Inference*, 2023c. URL https://openreview.net/forum?id=l-32qmxphio.

Vargas, F., Padhy, S., Blessing, D., and Nüsken, N. Transport meets variational inference: Controlled monte carlo diffusions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PP1rudnxiW.

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

Wainwright, M. J. and Jordan, M. I. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1:1–305, 2008. doi: 10.1561/2200000001. URL http://discovery.ucl.ac.uk/185880/.

Zhang, Q. and Chen, Y. Path Integral Sampler: a stochastic control approach for sampling. In *The Tenth International Conference on Learning Representations*, 2022.

Organization of the supplementary

The appendix is organized as follows. Appendix A summarizes general facts that will be useful for proofs. Appendix B provides details on the Stochastic Localization (SL) frameworks introduced in Section 2 and Section 3. In Appendix C, we derive detailed computations on the application of SL to Gaussian mixtures. Appendix D presents an implementation of SL for sampling in the case where the score of the observation process is available. Appendix E dispenses theoretical results on the phenomenon of "duality of log-concavity" highlighted in Section 4.2 and Section 4.3. In Appendix F, we detail the implementation of our sampling algorithm called `SLIPS` and provide an ablation study of its hyper-parameters. A detailed comparison between our approach and works from (Huang et al., 2024) and (Saremi et al., 2024) is given in Appendix G. Finally, Appendix H provides more details on the numerical experiments presented in Section 6.

Notation. To alleviate the computations derived below, for a probability distribution μ defined on $\mathbb{R}^d$, and any measurable function $\varphi : \mathbb{R}^d \to \mathbb{R}^d$, we will denote the expectation $\int_{\mathbb{R}^d} \varphi(x) \mathrm{d}\mu(x) \in \mathbb{R}^d$ by $\mathbb{E}[\varphi(X)]$ and the covariance $\int_{\mathbb{R}^d} (\varphi(x) - \mu(\varphi))(\varphi(x) - \mu(\varphi))^\top \mathrm{d}\mu(x) \in \mathbb{R}^{d\times d}$ by $\mathrm{Cov}[\varphi(X)]$, where X is a random vector distributed according to μ. For any $T \in (0, +\infty]$, two functions f and g defined on $[0, T)$ are said to be asymptotically equivalent if $f(t)/g(t) \to 1$ and $g(t)/f(t) \to 1$ as $t \to T$. This is denoted by $f(t) \sim g(t)$.

A. Preliminaries

Metrics. We recall that the 2-Wasserstein distance between two distributions μ and ν is given by

$$W_2(\mu, \nu) = \inf\{\textstyle\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x_1 - x_0\|^2 \mathrm{d}\pi(x_0, x_1) : \pi_0 = \mu, \pi_1 = \nu\}^{1/2} ,$$

where π_i denotes the i-th marginal of π for $i \in \{0, 1\}$. While it is not tractable in general, we have an explicit expression when comparing two Gaussian distributions. Consider $\mu = \mathsf{N}(\mathbf{m}_1, \gamma_1^2 \mathrm{I}_d)$ and $\nu = \mathsf{N}(\mathbf{m}_2, \gamma_2^2 \mathrm{I}_d)$, where $(\mathbf{m}_1, \mathbf{m}_2) \in \mathbb{R}^d \times \mathbb{R}^d$ and $(\gamma_1, \gamma_2) \in (0, \infty)^2$. Then, following (Peyré et al., 2019, Equation 2.41), we have

$$W_2(\mu, \nu)^2 = \|\mathbf{m}_1 - \mathbf{m}_2\|^2 + d(\gamma_1 - \gamma_2)^2 . \tag{16}$$

We also recall that the entropic regularized 2-Wasserstein distance between two distributions μ and ν is defined by

$$W_{2,\varepsilon}(\mu, \nu) = \inf\{\textstyle\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x_1 - x_0\|^2 \mathrm{d}\pi(x_0, x_1) - \mathscr{H}(\pi) : \pi_0 = \mu, \pi_1 = \nu\}^{1/2} ,$$

where $\varepsilon > 0$ is a regularization hyper-parameter and $\mathscr{H}(\pi) = -\int_{\mathbb{R}^d\times\mathbb{R}^d} \log \pi(x_0, x_1) \mathrm{d}\pi(x_0, x_1)$ refers to as the entropy of π.

Finally, we recall that the Kolmogorov-Smirnov distance between two distributions μ and ν is defined by

$$\mathrm{KS}(\mu, \nu) = \sup_{x\in\mathbb{R}^d} |F_\mu(x) - F_\nu(x)| ,$$

where F_μ (respectively F_ν) denotes the cumulative distribution function of μ (respectively ν).

Below, we provide a first result, known as the Tweedie's formula, which gives the expression of the score of any distribution that writes as a marginalization, and its derivative.

Lemma 4 (Tweedie's formula and extension)**.** *Let p be a positive probability density on $\mathbb{R}^d$ such that for any $y \in \mathbb{R}^d$ $p(y) = \int_{\mathbb{R}^d} p(y|x)q(x)\mathrm{d}x$, where $(x, y) \mapsto p(y|x)$ is a positive transition density on $\mathbb{R}^d \times \mathbb{R}^d$, and q is a positive density on $\mathbb{R}^d$. Assume that for any $x \in \mathbb{R}^d$, $y \mapsto \log p(y|x)$ is twice continuously differentiable and that, for any $k \in \{1, 2\}$, there exists $\varphi_k : \mathbb{R}^d \to \mathbb{R}_+$ such that $\int_{\mathbb{R}^d} \varphi_k(x)q(x)\mathrm{d}x < \infty$ and for any $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, we have $\left\|\nabla_y^k p(y|x)\right\| \le \varphi_k(x)$.*

For any $y \in \mathbb{R}^d$, define the 'posterior' density $x \mapsto p(x|y)$ by $p(x|y) = p(y|x)p(x)/p(y)$.

Then, we have for any $y \in \mathbb{R}^d$

$$\nabla_y \log p(y) = \int_{\mathbb{R}^d} \nabla_y \log p(y|x) p(x|y) \mathrm{d}x . \tag{17}$$

This is often referred to as the Tweedie's formula (Robbins, 1992). We also have for any $y \in \mathbb{R}^d$

$$\nabla_y^2 \log p(y) = \int_{\mathbb{R}^d} \nabla_y^2 \log p(y|x) p(x|y) \mathrm{d}x + \int_{\mathbb{R}^d} \nabla_y \log p(y|x) \{\nabla_y \log p(y|x)\}^\top p(x|y) \mathrm{d}x \\ - \left\{ \int_{\mathbb{R}^d} \nabla_y \log p(y|x) p(x|y) \mathrm{d}x \right\} \left\{ \int_{\mathbb{R}^d} \nabla_y \log p(y|x) p(x|y) \mathrm{d}x \right\}^\top . \tag{18}$$

For any $y \in \mathbb{R}^d$, these two results can be reformulated as

$$\nabla_y \log p(y) = \mathbb{E}[\nabla_y \log p(y|X)] \ , \ \nabla_y^2 \log p(y) = \mathbb{E}[\nabla_y^2 \log p(y|X)] + \mathrm{Cov}[\nabla_y \log p(y|X)] \ ,$$

where X is a random vector distributed according to the posterior distribution defined by $\mathrm{d}\mu^y(x) = p(x|y)\mathrm{d}x$.

Proof. First observe that for any $x \in \mathbb{R}^d$, $y \mapsto p(y|x)$ is twice continuously differentiable since $p(y|x) = \exp(\log p(y|x))$, where $y \mapsto \log p(y|x)$ is twice continuously differentiable.

We begin with the proof of (17). By combining the assumptions of the lemma with the dominated convergence theorem, we obtain that $p \in \mathrm{C}^2(\mathbb{R}^d, (0, \infty))$, and that for any $y \in \mathbb{R}^d$, we have

$$\nabla_y \log p(y) = \frac{\nabla_y p(y)}{p(y)} = \frac{\nabla_y \int_{\mathbb{R}^d} p(y|x) q(x) \mathrm{d}x}{\int_{\mathbb{R}^d} p(y|x) q(x) \mathrm{d}x} = \frac{\int_{\mathbb{R}^d} \nabla_y \log p(y|x) p(y|x) q(x) \mathrm{d}x}{\int_{\mathbb{R}^d} p(y|x) q(x) \mathrm{d}x} = \int_{\mathbb{R}^d} \nabla_y \log p(y|x) p(x|y) \mathrm{d}x \ ,$$

We now turn to the proof of (18). Note that we have $\nabla_y \log p(y) = p(y)^{-1} \int_{\mathbb{R}^d} \nabla_y \log p(y|x) p(y|x) q(x) \mathrm{d}x$. Then, using once again the dominated convergence theorem, we obtain that for any $y \in \mathbb{R}^d$,

$$\begin{aligned} \nabla_y^2 \log p(y) &= \frac{1}{p(y)} \int_{\mathbb{R}^d} \nabla_y^2 \log p(y|x) p(y|x) q(x) \mathrm{d}x + \frac{1}{p(y)} \int_{\mathbb{R}^d} \nabla_y \log p(y|x) \{\nabla_y \log p(y|x)\}^\top p(y|x) q(x) \mathrm{d}x \\ &\quad - \frac{\nabla_y p(y)}{p(y)^2} \left\{ \int_{\mathbb{R}^d} \nabla_y \log p(y|x) p(y|x) q(x) \mathrm{d}x \right\}^\top \\ &= \int_{\mathbb{R}^d} \nabla_y^2 \log p(y|x) p(x|y) \mathrm{d}x + \int_{\mathbb{R}^d} \nabla_y \log p(y|x) \{\nabla_y \log p(y|x)\}^\top p(x|y) \mathrm{d}x \\ &\quad - \nabla_y \log p(y) \left\{ \frac{1}{p(y)} \int_{\mathbb{R}^d} \nabla_y \log p(y|x) p(y|x) q(x) \mathrm{d}x \right\}^\top , \end{aligned}$$

which gives the result. □

We now dispense a useful lemma to compute exact integration in SDEs with linear drift.

Lemma 5. *Let $T > 0$. Consider the SDE defined on $[0, T]$ by $\mathrm{d}Y_t = \beta_t(Y_t + b)\mathrm{d}t + \sigma \mathrm{d}B_t$, where β is integrable on $[0, T]$, $b \in \mathbb{R}^d$, and $\sigma > 0$. For any $T \geq t > s \geq 0$, the conditional density of Y_t given $Y_s = y_s$, denoted by $p_{t|s}(\cdot|y_s)$, verifies*

$$p_{t|s}(y_t|y_s) = \mathsf{N}(y_t; \exp(\textstyle\int_s^t \beta_u \mathrm{d}u) y_s + (\exp(\int_s^t \beta_u \mathrm{d}u) - 1) b, \sigma^2 \int_s^t \exp(2 \int_u^t \beta_r \mathrm{d}r) \mathrm{d}u \, \mathrm{I}_d) \ .$$

Proof. Define $\gamma(t) = \exp(-\int_0^t \beta_u \mathrm{d}u)$. Consider the stochastic process $Z_t = \gamma(t) Y_t$. By Îto's formula, we have $\mathrm{d}Z_t = \beta_t \gamma(t) b \mathrm{d}t + \gamma(t) \sigma \mathrm{d}B_t = -\dot{\gamma}(t) b \mathrm{d}t + \gamma(t) \sigma \mathrm{d}B_t$. By integrating this SDE between s and t, we have

$$\gamma(t) Y_t - \gamma(s) Y_s = \{\gamma(s) - \gamma(t)\} b + \sigma \textstyle\int_s^t \gamma(u) \mathrm{d}B_u \ ,$$

and then

$$Y_t = \exp(\textstyle\int_s^t \beta_u \mathrm{d}u) Y_s + (\exp(\int_s^t \beta_u \mathrm{d}u) - 1) b + \sigma \int_s^t \exp(\int_u^t \beta_r \mathrm{d}r) \mathrm{d}B_u \ ,$$

which gives the result using Îto's isometry and the fact that Y_s is independent from $(B_t - B_s)_{t \in [s,T]}$. □

B. Details on Stochastic Localization and our extension

Vocabulary of Stochastic Localization (SL). For sake of clarity, we precise below some terms employed in our paper:

- Y_t^α: observation process,
- g: denoising schedule,
- u_t^α : denoiser function,
- $u_t^\alpha(Y_t^\alpha)$: denoiser or Bayes estimator,
- $q_t^\alpha(\cdot|Y_t^\alpha)$: random posterior density. In the geometric measure theory literature, this is referred to as the SL process.

B.1. Reminders and intuition on generalized stochastic localization

Connection between standard stochastic localization and diffusion models. Stochastic localization as defined in (1) is equivalent to Variance-Preserving (VP) diffusion models (Song et al., 2021) under change of time. Indeed, if we define $\Phi(t) = \log(1 + \sigma^2 t^{-1})/2$ for any $t \geq 0$, with the convention that $\Phi(0) = \infty$, and consider the Ornstein-Uhlenbeck process $(X_t)_{t\geq 0}$ solving the SDE

$$\mathrm{d}X_s = -X_s \mathrm{d}s + \sqrt{2}\mathrm{d}B_s \ , \quad X_0 \sim \pi \ , \tag{19}$$

then, $(Y_t)_{t\geq 0}$ and $(\sqrt{t(t+\sigma^2)}X_{\Phi(t)})_{t\geq 0}$ have the same distribution by application of Dubins-Schwartz theorem (Montanari, 2023). In other words, the localization process (1) can be identified as a *non-linear* time-reversal transformation of the noising VP process (19).

Comments on our framework. We recall that $\alpha(t) = t^{1/2}g(t)$. We now justify the assumptions given on g in Section 3.1.

- Rationale behind (a) and (b): together, these requirements ensure that (i) α is continuously derivable at $t = 0$, without imposing this same condition on g and that (ii) we have $\alpha(0) = 0$. With (i), the SDE defined in (6) does not have a singularity at time $t = 0$. With (ii), it allows to fix Y_0^α to a deterministic value. This is equivalent to full independence with X, *i.e.*, full noise looking at the expression of the SNR in (11), which is natural to start the denoising process. Note that taking $-g$ instead of g would be equivalent in the denoising procedure, since the SNR considers $g(t)^2$. Nonetheless, considering g with positive values is arbitrary and allows us to simplify our framework.
- Rationale behind (c): as $t \to T_{\text{gen}}$, we would like to denoise increasingly the observation process, *i.e.*, gaining progressively information about X, such that we obtain complete denoising at the end of the process. Looking at the expression of the SNR in (11), this requires to naturally take $g(t) \to \infty$ as $t \to T_{\text{gen}}$ and g strictly increasing, recalling that g takes non negative values.

Note that it includes the case of standard stochastic localization by taking $g(t) = t^{1/2}$. In this case, all properties are verified: in particular, consider $C = 1$ and $\beta = 1$ in (b).

B.2. Theoretical results.

Here, we provide details on the theoretical claims made in Section 3.1 on our extended framework of stochastic localization.

Link between score and denoiser function. The formula given in (9) is an immediate corollary of Lemma 4 applied to the observation process defined in (4). Indeed, the conditional distribution of Y_t^α given $X = x \in \mathbb{R}^d$ is $p_t^\alpha(\cdot|x) = \mathsf{N}(\alpha(t)x, \sigma^2 t\, \mathrm{I}_d)$. Then, it comes that $\nabla_y \log p_t^\alpha(y|x) = \{\alpha(t)x - y\}/\sigma^2 t$, and we obtain the result using Lemma 4-(17).

Localization rate. We begin with the localization rate given in Equation (5). We define the following quantity $\mathrm{LocR}(t) = \sigma\sqrt{d}/g(t)$. We emphasize that the assumption of finite second order moment on the target distribution, that is used below, guarantees the proper definition of the 2-Wasserstein distance.

Proposition 6. *Assume that π has finite second order moment. Denote by π_t^α the probability distribution of $Y_t^\alpha/\alpha(t)$ where $(Y_t^\alpha)_{t\in[0,T_{gen})}$ is the stochastic observation process defined in (4). Then, for any $t \in (0, T_{gen})$, we have*

$$W_2(\pi, \pi_t^\alpha) \leq \mathrm{LocR}(t) \ .$$

Proof. Let $t \in (0, T_{\text{gen}})$. Consider the following coupling $(X, \hat{Y}_t^\alpha)$: (i) $X \sim \pi$, (ii) $\hat{Y}_t^\alpha = \alpha(t)X + \sigma\sqrt{t}Z$, where $Z \sim \mathsf{N}(0, \mathrm{I}_d)$. Denote $X_t^\alpha = \hat{Y}_t^\alpha/\alpha(t)$. We have $X_t^\alpha = X + \{\sigma/g(t)\}\, Z$ and $\hat{X}_t^\alpha \sim \pi_t^\alpha$. Therefore, it comes that

$$W_2(\pi, \pi_t^\alpha)^2 \leq \mathbb{E}[\|\sigma/g(t)\, Z\|^2] = \sigma^2 d/g(t)^2 \ ,$$

which gives the result. □

In particular, we recover the localization rate of the standard setting given in Section 2 with $g(t) = t^{1/2}$ and $T_{\text{gen}} = \infty$. Note that this upper bound applies on general distributions that may be non log-concave, as considered in **A**0, and may be improved by assuming further regularity. We notably show a strong refinement in the Gaussian case by a factor $O(g(t))$.

Proposition 7. *Consider the target distribution given by $\pi = \mathsf{N}(\mathbf{m}, \gamma^2 \mathrm{I}_d)$ where $\mathbf{m} \in \mathbb{R}^d$ and $\gamma > 0$. Then, as $t \to T_{gen}$, we have*

$$W_2(\pi, \pi_t^\alpha) = \gamma \left| 1 - \left(1 + \frac{\sigma^2}{\gamma^2 g(t)^2} \right)^{1/2} \right| \sqrt{d} \sim \frac{\sigma}{2\gamma g(t)} \mathrm{LocR}(t) \ .$$

Proof. In this setting, the distribution of the observation process at time t is tractable, and we have

$$p_t^\alpha = \mathsf{N}(\alpha(t)\mathbf{m}, \{\alpha(t)^2\gamma^2 + \sigma^2 t\}\ \mathrm{I}_d) \ .$$

Then, we get that

$$\pi_t^\alpha = \mathsf{N}(\mathbf{m}, \{\gamma^2 + \sigma^2/g(t)^2\}\, \mathrm{I}_d) \ .$$

In this particular, using (16), the 2-Wasserstein distance is given by

$$W_2(\pi, \pi_t^\alpha) = \gamma \left| 1 - \left(1 + \frac{\sigma^2}{\gamma^2 g(t)^2} \right)^{1/2} \right| \sqrt{d} \ ,$$

from which we deduce the result by a simple asymptotic equivalent. □

Interestingly, the same bound applies on the denoiser $u_t^\alpha(Y_t^\alpha) = \int_{\mathbb{R}^d} x q_t^\alpha(x|Y_t^\alpha)$. The proof relies on the same structure.

Proposition 8. *Assume that π has finite second order moment. Denote by $\tilde{\pi}_t^\alpha$ the probability distribution of $u_t^\alpha(Y_t^\alpha)$ where $(Y_t^\alpha)_{t\in[0,T_{gen})}$ is the stochastic observation process defined in (4). Then, for any $t \in (0, T_{gen})$, we have*

$$W_2(\pi, \tilde{\pi}_t^\alpha) \leq \mathrm{LocR}(t) \ .$$

Proof. Let $t \in (0, T_{\text{gen}})$. Consider the following coupling $(X, \hat{Y}_t^\alpha)$: (i) $X \sim \pi$, (ii) $\hat{Y}_t^\alpha = \alpha(t)X + \sigma\sqrt{t}Z$, where $Z \sim \mathsf{N}(0, \mathrm{I}_d)$. Denote $X_t^\alpha = u_t^\alpha(Y_t^\alpha)$. We have $X_t^\alpha \sim \tilde{\pi}_t^\alpha$. Therefore, it comes that

$$W_2(\pi, \tilde{\pi}_t^\alpha)^2 \leq \mathbb{E}_{X,X_t^\alpha}[\|X_t^\alpha - X\|^2] = \mathbb{E}_{X,\hat{Y}_t^\alpha}[\|\mathbb{E}[X|\hat{Y}_t^\alpha] - X\|^2] = \mathbb{E}_{X,\hat{Y}_t^\alpha}[\|\mathbb{E}[X|\hat{Y}_t^\alpha/\alpha(t)] - X\|^2] \ .$$

Since conditional expectations are orthogonal projections in L^2, we have

$$W_2(\pi, \tilde{\pi}_t^\alpha)^2 \leq \mathbb{E}_{X,\hat{Y}_t^\alpha}[\|\hat{Y}_t^\alpha/\alpha(t) - X\|^2] = \mathbb{E}[\|\sigma/g(t)\, Z\|^2] = \sigma^2 d/g(t)^2 \ ,$$

which gives the result. □

Similarly, we also show a strong refinement of Proposition 8 by a factor $O(g(t))$ in the Gaussian case.

Proposition 9. *Consider the target distribution given by $\pi = \mathsf{N}(\mathbf{m}, \gamma^2 \mathrm{I}_d)$ where $\mathbf{m} \in \mathbb{R}^d$ and $\gamma > 0$. Then, as $t \to T_{gen}$, we have*

$$W_2(\pi, \tilde{\pi}_t^\alpha) = \gamma \left| 1 - \left(1 + \frac{\sigma^2}{\gamma^2 g(t)^2} \right)^{-1/2} \right| \sqrt{d} \sim \frac{\sigma}{2\gamma g(t)} \operatorname{LocR}(t) \, .$$

Proof. Here again, the distribution of the observation process at time t is tractable, we have

$$p_t^\alpha = \mathsf{N}(\alpha(t)\mathbf{m}, \{\alpha(t)^2\gamma^2 + \sigma^2 t\} \, \mathrm{I}_d) \, . \tag{20}$$

In particular, the score is tractable and given by

$$\nabla_y \log p_t^\alpha(y) = \frac{\alpha(t)\mathbf{m} - y}{\alpha(t)^2\gamma^2 + \sigma^2 t} \, .$$

Using relation (9), we thus have

$$u_t^\alpha(y) = \frac{\alpha(t)\gamma^2}{\alpha(t)^2\gamma^2 + \sigma^2 t} y + \frac{\sigma^2 t}{\alpha(t)^2\gamma^2 + \sigma^2 t} \mathbf{m} \, . \tag{21}$$

Recalling that $\tilde{\pi}_t^\alpha$ is the distribution of $u_t^\alpha(Y_t^\alpha)$, by combining (20) and (21), it comes that

$$\tilde{\pi}_t^\alpha = \mathsf{N}\left(\mathbf{m}, \gamma^2 \left\{ 1 + \tfrac{\sigma^2}{\gamma^2 g(t)^2} \right\}^{-1} \mathrm{I}_d \right) \, .$$

In this particular, using (16), the 2-Wasserstein distance is given by

$$W_2(\pi, \tilde{\pi}_t^\alpha) = \gamma \left| 1 - \left(1 + \frac{\sigma^2}{\gamma^2 g(t)^2} \right)^{-1/2} \right| \sqrt{d}$$

from which we deduce the result with a simple asymptotic equivalent. □

General remark on denoising in Stochastic Localization. In our presentation of standard stochastic localization where the observation process is defined in (1), we take Y_T/T as approximate sample from π, where Y_T is obtained by running the (discretized) SDE (2) up to time T. Yet, in standard stochastic localization, (Montanari, 2023) proposed to take the last denoiser of the procedure $u_T(Y_T) = \int_{\mathbb{R}^d} x q_T(x|Y_T) \mathrm{d}x$. We highlight here that this is equivalent asymptotically.

- The two approaches have the same upper bound on their rate of convergence, see Proposition 6 and Proposition 8 in the standard setting. Note that we also obtain the same rate asymptotically in the Gaussian case, see Proposition 7 and Proposition 9.
- When T is large, the Gaussian term dominates the target term in the expression of the posterior q_t given in (3). Noting that $\sigma^2/T \ll 1$, it comes that $q_T(x|Y_T) \approx \delta_{Y_T/T}$, and therefore $u_T(Y_T) \approx Y_T/T$.

As shown above, this equivalence also occurs in generalized stochastic localization. In addition to the similarity of results in Proposition 6 and Proposition 8, we observe that, as T is large if $T_{\text{gen}} = \infty$ or as T is close to T_{gen} otherwise, the Gaussian term dominates the target term in the expression of the posterior q_t given in (8). Since $\sigma^2/g(T)^2 \ll 1$, it comes that $q_T^\alpha(x|Y_T^\alpha) \approx \delta_{Y_T^\alpha/\alpha(T)}$, and therefore, we have again $u_T^\alpha(Y_T^\alpha) \approx Y_T^\alpha/\alpha(T)$.

In the main document of our paper, we choose to consider the distribution of $Y_T^\alpha/\alpha(T)$ rather than the the distribution of the denoiser $u_T^\alpha(Y_T^\alpha)$ as an approximation of the target distribution. This choice seemed to us to be easier to understand for non-expert readers. In practice, we however align with the original methodology and compute the denoiser $u_T^\alpha(Y_T^\alpha)$ instead of $Y_T^\alpha/\alpha(T)$. This allows to us to have a fair comparison with the approach from (Saremi et al., 2024), who also consider such Bayes estimator for sampling.

Markovian projection of the observation process. Following (Liptser & Shiryaev, 1977), we now turn to existence and uniqueness results on the SDE (6), under assumptions on π. These are corollaries of the general results: (Liptser & Shiryaev, 1977, Theorem 7.12), restated in Proposition 10, and (Liptser & Shiryaev, 1977, Theorem 7.6), restated in Proposition 12.

Proposition 10 (Theorem 7.12 in (Liptser & Shiryaev, 1977))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space on $\mathbb{R}^d$, let $(\mathcal{F}_t)_{t\geq 0}$ be a nondecreasing family of sub-σ-algebras and let $W = (W_t, \mathcal{F}_t)_{t\geq 0}$ be a Wiener process. Let $T > 0$, $\sigma > 0$. Consider the Îto process $Y = (Y_t, \mathcal{F}_t)_{t\in[0,T]}$ satisfying the following SDE*

$$\mathrm{d}Y_t = \beta_t \mathrm{d}t + \sigma \mathrm{d}W_t, \quad Y_0 = 0,$$

where the stochastic process $\beta = (\beta_t, \mathcal{F}_t)_{t\in[0,T]}$ verifies $\mathbb{P}(\int_0^T \|\beta_t\| \,\mathrm{d}t < \infty) = 1$.

Assume that $\int_0^T \mathbb{E}[\|\beta_t\|]\mathrm{d}t < \infty$ and define $h : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ by

$$h_t(y) = \mathbb{E}[\beta_t | Y_t = y] \,.$$

Also define the stochastic process $(\overline{W}_t)_{t\in[0,T]}$ as

$$\overline{W}_t = \frac{1}{\sigma}\left(Y_t - \int_0^t h_s(Y_s)\mathrm{d}s\right) \,.$$

Then, $\overline{W}$ is a Wiener process and Y is a diffusion process satisfying the following SDE

$$\mathrm{d}Y_t = h_t(Y_t)\mathrm{d}t + \sigma \mathrm{d}\overline{W}_t \,, Y_0 = 0 \,.$$

Since the original result considers the case where $\sigma = 1$ and $d = 1$, we provide below the proof of Proposition 10 for completeness.

Proof. Let $t \in [0,T]$. We have that $Y_t = Y_0 + \int_0^t \beta_s \mathrm{d}s + \sigma \int_0^t \mathrm{d}W_s = \int_0^t \beta_s \mathrm{d}s + \sigma W_t$. Hence, we get that

$$\overline{W}_t = \frac{1}{\sigma}\int_0^t [\beta_s - h_s(Y_s)]ds + W_t \,.$$

In particular, $\overline{W}_0 = 0$. Using Itô's formula, we have that

$$\mathrm{d}(\exp(iz^\top \overline{W}_t)) = i\exp(iz^\top \overline{W}_t) z^\top \mathrm{d}\overline{W}_t - \frac{\|z\|^2}{2}\exp(iz^\top \overline{W}_t)\mathrm{d}t \,.$$

Integrating this formula between s and t, where $0 \leq s \leq t \leq T$, we obtain that

$$\begin{aligned}\exp(iz^\top(\overline{W}_t - \overline{W}_s)) = 1 &+ iz^\top \int_s^t \exp(iz^\top(\overline{W}_u - \overline{W}_s))\mathrm{d}W_u \\ &+ \frac{iz^\top}{\sigma}\int_s^t \exp(iz^\top(\overline{W}_u - \overline{W}_s))[\beta_u - h_u(Y_u)]\mathrm{d}u \\ &- \frac{\|z\|^2}{2}\int_s^t \exp(iz^\top(\overline{W}_u - \overline{W}_s))\mathrm{d}u \,.\end{aligned}$$

Using $\mathbb{E}[\int_s^t \exp(iz^\top(\overline{W}_u - \overline{W}_s))\mathrm{d}W_u | Y_s] = 0$ and

$$\mathbb{E}\left[\int_s^t \exp(iz^\top(\overline{W}_u - \overline{W}_s))[\beta_u - h_u(Y_u)]\mathrm{d}u | Y_s\right] = \mathbb{E}\left[\int_s^t \exp(iz^\top(\overline{W}_u - \overline{W}_s))\mathbb{E}[\beta_u - h_u(Y_u)|Y_u]\mathrm{d}u | Y_s\right] = 0 \,,$$

by definition of h_u, we get that

$$\mathbb{E}[\exp(iz^\top(\overline{W}_t - \overline{W}_s))|Y_s] = 1 - \frac{\|z\|^2}{2}\int_s^t \mathbb{E}[\exp(iz^\top(\overline{W}_u - \overline{W}_s))|Y_s]\mathrm{d}u \,.$$

Define $f(t) = \mathbb{E}[\exp(iz^\top(\overline{W}_t - \overline{W}_s))|Y_s]$. Since $f(t) = 1 - \frac{\|z\|^2}{2}\int_s^t f(u)\mathrm{d}u$, f is continuously differentiable on $[s,t]$. Following our computations, we have $\dot{f}(t) = -(\|z\|^2/2)f(t)$ with $f(s) = 1$, which implies that $f(t) = \exp(-\frac{\|z\|^2}{2}(t-s))$. Consequently, $\overline{W}$ is a Wiener process with the filtration $(\mathcal{F}_t)_{t\in[0,T]}$. By definition of $\overline{W}$, this means that Y satisfies the following SDE

$$\mathrm{d}Y_t = h_t(Y_t)\mathrm{d}t + \sigma\mathrm{d}\overline{W}_t \ , Y_0 = 0 \ .$$

□

Corollary 11. *Assume that π has finite first order moment. Consider the observation process Y^α defined in (4). Then it is a solution to the SDE*

$$\mathrm{d}Y_t^\alpha = \dot{\alpha}(t)u_t^\alpha(Y_t^\alpha)\mathrm{d}t + \sigma\mathrm{d}B_t, \ Y_0^\alpha = 0 \ .$$

Proof. Consider $T \in (0, T_{\text{gen}})$. We recall that $Y_t^\alpha = \alpha(t)X + \sigma W_t$. Then, $(Y_t^\alpha)_{t\in[0,T]}$ is an Îto process satisfying the SDE

$$\mathrm{d}Y_t^\alpha = \beta_t\mathrm{d}t + \sigma\mathrm{d}B_t, \ Y_0^\alpha = 0 \ ,$$

where $\beta_t = \dot{\alpha}(t)X$. Note that β is well defined on $[0,T]$ due to the assumptions (a) and (b) considered on g and that we immediately have $\mathbb{P}(\int_0^T \|\beta_t\| \, \mathrm{d}t < \infty) = 1$. Moreover, since α is strictly increasing on $[0,T]$, $|\dot{\alpha}| = \dot{\alpha}$ and we get that

$$\int_0^T \mathbb{E}[\|\beta_t\|]\mathrm{d}t = \mathbb{E}[\|X\|]\int_0^T \dot{\alpha}(t)\mathrm{d}t = \mathbb{E}[\|X\|]\{\alpha(T) - \alpha(0)\} \ .$$

Since $\mathbb{E}[\|X\|] < \infty$ by assumption on π, we thus have $\int_0^T \mathbb{E}[\|\beta_t\|]\mathrm{d}t < \infty$. By defining $h_t(y) = \int_{\mathbb{R}^d} \dot{\alpha}(t)xq_t^\alpha(x|y)\mathrm{d}x = \dot{\alpha}(t)u_t^\alpha(y)$, we finally obtain the result by applying Proposition 10. □

Proposition 12 (Theorem 7.6 & Remark 7.2.7. in (Liptser & Shiryaev, 1977))**.** *Let $T > 0$, $\sigma > 0$. Consider the following SDE*

$$\mathrm{d}Y_t = h_t(Y_t)\mathrm{d}t + \sigma\mathrm{d}W_t, \quad Y_0 = 0 \ , \tag{22}$$

where $\mathbb{P}(\int_0^T \|h_t(Y_t)\|^2 \, \mathrm{d}t < \infty) = 1$. Then, (22) admits at most one weak solution.

The proof of this result lies in the fact that the distributions of solutions to the SDE (22) have the same density with respect to the distribution of the Brownian motion. We refer to (Liptser & Shiryaev, 1977) for the complete proof.

Corollary 13. *Assume that π has finite second order moment and that $t \to \dot{\alpha}(t)^2$ is integrable at time $t = 0$. Then, the SDE defined in (6) by*

$$\mathrm{d}Y_t^\alpha = \dot{\alpha}(t)u_t^\alpha(Y_t^\alpha)\mathrm{d}t + \sigma\mathrm{d}B_t \ , \ Y_0^\alpha = 0 \ ,$$

has a unique weak solution.

Proof. Consider $T \in (0, T_{\text{gen}})$. Since π has its first order moment that is finite, we have existence of solutions by Corollary 11. Consider a solution Y^α and denote $h_t = \dot{\alpha}(t)u_t^\alpha$. We have $\|h_t(Y_t^\alpha)\|^2 = \dot{\alpha}(t)^2\|\mathbb{E}[X|Y_t^\alpha]\|^2 \leq \dot{\alpha}(t)^2\mathbb{E}[\|X\|^2|Y_t^\alpha]$ by Jensen's inequality. Then, using this inequality, we obtain that

$$\mathbb{E}[\int_0^T \|h_t(Y_t^\alpha)\|^2 \, \mathrm{d}t] = \int_0^T \mathbb{E}[\|h_t(Y_t^\alpha)\|^2]\mathrm{d}t \leq \int_0^T \dot{\alpha}(t)^2\mathbb{E}[\mathbb{E}[\|X\|^2|Y_t^\alpha]]\mathrm{d}t = \mathbb{E}[\|X\|^2]\int_0^T \dot{\alpha}(t)^2\mathrm{d}t$$

Combining the assumptions on π and α, it comes that $\mathbb{E}[\int_0^T \|h_t(Y_t^\alpha)\|^2 \, \mathrm{d}t] < \infty$ using the inequality above, and therefore $\mathbb{P}(\int_0^T \|h_t(Y_t^\alpha)\|^2 \, \mathrm{d}t < \infty) = 1$. We finally obtain the result by applying Proposition 12. □

We emphasize that the extra assumption made on α in Corollary 13 is verified for the localization schemes Geom and Geom-∞ presented in Section 3.2.

B.3. Alternative denoising approach to SLIPS

We note here the score of the observation process can be expressed via an expectation over the posterior of the model, in a different manner than (9). Define $v_t^\alpha(y) = \int_{\mathbb{R}^d} \nabla_x \log \pi(x) q_t^\alpha(x|y) \mathrm{d}x$, where q_t^α is the posterior density given in (8).

Lemma 14. *Consider the observation process defined in (4) with marginal distribution at time t given by p_t^α. Assume that $\log \pi$ is continuously differentiable on $\mathbb{R}^d$ and that there exists $\varphi : \mathbb{R}^d \to \mathbb{R}_+$ such that $\int_{\mathbb{R}^d} \varphi(z) \mathsf{N}(z; 0, \sigma^2 t \mathrm{I}_d) \mathrm{d}z < \infty$ and for any $(z, y) \in \mathbb{R}^d \times \mathbb{R}^d$, we have $\|\nabla_y \pi(\{z+y\}/\alpha(t))\| \leq \varphi(z)$. Then, we have for any $y \in \mathbb{R}^d$*

$$\nabla_y \log p_t^\alpha(y) = v_t^\alpha(y)/\alpha(t) \ .$$

Proof. We recall that $p_t^\alpha(y) = \int_{\mathbb{R}^d} \mathsf{N}(y; \alpha(t)x, \sigma^2 t \, \mathrm{I}_d) \mathrm{d}\pi(x)$. Then, by change of variable $z = \alpha(t)x - y$, we have

$$p_t^\alpha(y) \propto \int_{\mathbb{R}^d} \pi(\{z+y\}/\alpha(t)) \mathsf{N}(z; 0, \sigma^2 t \, \mathrm{I}_d) \mathrm{d}z \ ,$$

where the multiplicative constant does not depend on y. For any $y \in \mathbb{R}^d$, denote by $\tilde{q}_t^\alpha(\cdot|y)$ the density defined up to a normalizing constant by $\tilde{q}_t^\alpha(z|y) \propto \pi(\{z+y\}/\alpha(t)) \mathsf{N}(z; 0, \sigma^2 t \, \mathrm{I}_d)$. By combining the assumptions of the lemma with the result of Lemma 4-(17) for this new expression of p_t^α, we obtain that

$$\begin{aligned}
\nabla_y \log p_t^\alpha(y) &= \int_{\mathbb{R}^d} \nabla_y \log \pi(\{z+y\}/\alpha(t)) \tilde{q}_t^\alpha(z|y) \mathrm{d}z \\
&= \frac{1}{\alpha(t)} \int_{\mathbb{R}^d} \nabla \log \pi(\{z+y\}/\alpha(t)) \tilde{q}_t^\alpha(z|y) \mathrm{d}z \\
&= \frac{1}{\alpha(t)} \int_{\mathbb{R}^d} \nabla_x \log \pi(x) q_t^\alpha(x|y) \mathrm{d}x \ ,
\end{aligned}$$

where we re-applied the change of variable $z = \alpha(t)x - y$ in the last equality. □

Therefore, the SDE (6) is strictly equivalent to the SDE

$$\mathrm{d}Y_t^\alpha = \tfrac{\dot{\alpha}(t)}{\alpha(t)} \{ Y_t^\alpha + \tfrac{\sigma^2 t}{\alpha(t)} v_t^\alpha(Y_t^\alpha) \} \mathrm{d}t + \sigma \mathrm{d}B_t \ , \ Y_0^\alpha = 0 \ .$$

Hence, to simulate from the observation process in a learning-free fashion, we can adopt a similar strategy to SLIPS, that rather involves the SDE derived above. Given a SNR-adapted discretization $(t_k)_{k=1}^K$ of a time interval $[t_0, T]$, where $t_0 > 0$ and $K \geq 1$, it amounts to consider a sequence $\{(\tilde{Y}_{t_k}^\alpha, V_{t_k}^\alpha)\}_{k=1}^K$, where $V_{t_k}^\alpha$ is a Monte Carlo estimator of $v_{t_k}^\alpha(\tilde{Y}_{t_k}^\alpha)$ and $\tilde{Y}_{t_k}^\alpha$ is obtained by solving the SDE

$$\mathrm{d}\tilde{Y}_t^\alpha = \tfrac{\dot{\alpha}(t)}{\alpha(t)} \{ \tilde{Y}_t^\alpha + \tfrac{\sigma^2 t_k}{\alpha(t_k)} V_{t_k} \} \mathrm{d}t + \sigma \mathrm{d}B_t \ , t \in [t_k, t_{k+1}] \ . \tag{23}$$

In theory, this approach leads to the same level difficulty in the Monte Carlo estimation as SLIPS, since the involved random poster densities are the same. In particular, this formulation does not bypass the "duality of log-concavity" explained in Section 4.3. The main difference with the approach presented in the main of our paper however lies in the integration of the SDE (23), which is not tractable at first sight. By combining the result of Lemma 5 and the use of the stochastic Exponential Integrator scheme (Durmus & Moulines, 2015), we show in Appendix D.1 how to solve this SDE for our localization schemes. We emphasize that this implementation is also available in our code.

Unfortunately, we found in our early experiments that this approach suffered from numerical unstability and showed higher variance than SLIPS. We think that this is due to the evaluation of $\nabla \log \pi$ in our MC estimation for the early steps of the algorithm, and could be overcome by using the SLIPS recursion given in (14) instead. We leave this study for future work.

C. Detailed computations for Gaussian mixtures

In this section, we consider the special case where π is a mixture of N Gaussians with weights $(w_i)_{i=1}^N$, means $(\mathbf{m}_i)_{i=1}^N$ and covariance matrices $(\gamma_i^2 \, \mathrm{I}_d)_{i=1}^N$. Under this assumption, p_t^α can be explicitly written for any $y \in \mathbb{R}^d$ and $t \in (0, T_{\text{gen}})$ as

$$p_t^\alpha(y) = \sum_{i=1}^N w_i \mathsf{N}(y; \alpha(t)\mathbf{m}_i, t(g^2(t)\gamma_i^2 + \sigma^2)\,\mathrm{I}_d) \; .$$

This means that the distribution of the observation process is itself a mixture of Gaussians with the same weights as π but with means $(\alpha(t)\mathbf{m}_i)_{i=1}^N$ and covariance matrices $(t(g^2(t)\gamma_i^2 + \sigma^2)\,\mathrm{I}_d)_{i=1}^N$. This elementary result is obtained by applying the rule of linear combination of independent Gaussian random variables. The score of the observation process can also be computed by noticing that for all $y \in \mathbb{R}^d$

$$\nabla \log p_t^\alpha(y) = \frac{\nabla p_t^\alpha(y)}{p_t^\alpha(y)} = -\frac{\sum_{i=1}^N w_i t^{-1}(g^2(t)\gamma_i^2 + \sigma^2)^{-1}(y - \alpha(t)\mathbf{m}_i)\mathsf{N}(y; \alpha(t)\mathbf{m}_i, t(g^2(t)\gamma_i^2 + \sigma^2)\,\mathrm{I}_d)}{\sum_{i=1}^N w_i \mathsf{N}(y; \alpha(t)\mathbf{m}_i(g^2(t)\gamma_i^2 + \sigma^2)\,\mathrm{I}_d)} \; .$$

One can also compute the posterior distribution q_t^α for any $t \in (0, T_{\text{gen}})$ and $x, y \in \mathbb{R}^d$ as

$$\begin{aligned}
q_t^\alpha(x|y) &\propto \sum_{i=1}^N w_i \mathsf{N}(x; \mathbf{m}_i, \gamma_i^2 \, \mathrm{I}_d)\mathsf{N}(y; \sqrt{t}g(t)x, \sigma^2 t \, \mathrm{I}_d) \\
&= \sum_{i=1}^N w_i (\sqrt{t}g(t))^{-d} \mathsf{N}(x; \mathbf{m}_i, \gamma_i^2 \, \mathrm{I}_d)\mathsf{N}\left(x; \frac{y}{\sqrt{t}g(t)}, \frac{\sigma^2}{g^2(t)}\mathrm{I}_d\right) \\
&= \sum_{i=1}^N \underbrace{w_i \alpha(t)^{-d} \mathsf{N}\left(\mathbf{m}_i; \frac{y}{\alpha(t)}, \left(\gamma_i^2 + \frac{\sigma^2}{g^2(t)}\right)\mathrm{I}_d\right)}_{=\tilde{w}_{t,y,i}^\alpha} \mathsf{N}\left(x; \underbrace{\left(\frac{\gamma_i^2\sigma^2}{\sigma^2 + g^2(t)\gamma_i^2}\right)\left(\frac{\mathbf{m}_i}{\gamma_i^2} + \frac{yg(t)}{\sqrt{t}\sigma^2}\right)}_{=\mathbf{m}_{t,y,i}^\alpha}, \underbrace{\left(\frac{\gamma_i^2\sigma^2}{\sigma^2 + g^2(t)\gamma_i^2}\right)}_{=(\gamma_{t,y,i}^\alpha)^2}\mathrm{I}_d\right) \; .
\end{aligned}$$

This shows that the posterior is itself a mixture of Gaussian distributions with weights $(w_{t,y,i}^\alpha)_{i=1}^N$ where $w_{t,y,i}^\alpha = \tilde{w}_{t,y,i}^\alpha / \sum_{j=1}^N \tilde{w}_{t,y,j}^\alpha$, means $(\mathbf{m}_{t,y,i}^\alpha)_{i=1}^N$ and covariance matrices $((\gamma_{t,y,i}^\alpha)^2 \, \mathrm{I}_d)_{i=1}^N$.

Additionally, we can derive tight expressions for the constants R and τ introduced in **A**0 in the case where π is a Gaussian mixture parameterized by $N = 2$, $w_1 = 1 - w_2 = w$, with $w \in (0, 1)$, $\mathbf{m}_2 = -\mathbf{m}_1 = a \, \mathbf{1}_d$[7], with $a > 0$, and $\gamma_1 = \gamma_2 = \gamma$, with $\gamma > 0$. In this case, π verifies **A**0, where (i) μ is a mixture of two Dirac masses at $-a \, \mathbf{1}_d$ and $+a \, \mathbf{1}_d$ with respective weights w and $1 - w$ and (ii) $\tau = \gamma$. Moreover, for any random vector $U \sim \mu$, it holds almost surely that

$$\|U - \mathbf{m}_\pi\| = \|U - a(1 - 2w) \, \mathbf{1}_d\| \leq 2 \max(w, 1 - w) a \sqrt{d} \; .$$

Therefore, we obtain that $R = 2 \max(w, 1 - w)a$ in **A**0. Note also that $\bar{\pi}$, defined as the distribution of $X - \mathbf{m}_\pi$ where $X \sim \pi$, is still a Gaussian mixture that verifies **A**0 with same constants R and τ.

In the experiments conducted in Section 6, Appendix D and Appendix F.2, we will consider a rolling example given by the target distribution $\bar{\pi}$ where $w = 1/3$, $a = 1.0$ and $\gamma^2 = 0.05$. In this case, we have $R = 4/3$ and $\tau^2 = 0.05$ and the corresponding density is defined as

$$x \in \mathbb{R}^d \mapsto \tfrac{2}{3}\mathsf{N}\left(x; -\tfrac{2}{3} \, \mathbf{1}_d, 0.05 \, \mathrm{I}_d\right) + \tfrac{1}{3}\mathsf{N}\left(x; \tfrac{4}{3} \, \mathbf{1}_d, 0.05 \, \mathrm{I}_d\right) \; . \tag{24}$$

We recall that $d(R^2 + \tau^2)$ is an upper bound on the scalar variance of this target distribution.

[7] We recall that $\mathbf{1}_d$ stands for the d-dimensional vector with all components equal to 1.

D. Sampling via Stochastic Localization in an ideal setting

The goal of this section is to validate the claims from Section 4.1 about the minimization of the integration error. In this section, we work under the assumption that the score $\nabla \log p_t^\alpha$ is a known function and do not consider any MCMC method at all (even for the initialization). This setting removes entirely the estimation error that we deal with in Section 4.2 to solely focus on the integration error. For the initialization, we now consider an arbitrary $t_0 > 0$ with $Y_{t_0}^\alpha$ distributed as $\mathsf{N}(0, \sigma^2 t_0 \, \mathrm{I}_d)$ (the best estimation that we have at the beginning of the SDE). We will clarify 4 different points in this section.

(a) Exploring *Exponential Integration* (EI) based schemes for SDE discretization;

(b) Analyzing the impact of the SNR-adapted discretization on the integration error;

(c) Exploring the limits of the Gaussian approximation $Y_{t_0}^\alpha \sim \mathsf{N}(0, \sigma^2 t_0 \, \mathrm{I}_d)$;

(d) Exploring the impact of the computational budget.

In the numerical examples presented below, we consider the target distribution defined in (24) where $d = 10$. We compute the exact score using the analytical formulas from Appendix C. We choose to display two complementary results based on (i) the empirical Sliced Wasserstein distance (Bonneel et al., 2015; Nadjahi et al., 2019), which tells how local information on π is recovered, and (ii) the error in estimating the weight of the first mode, which tells about global properties of the estimation. Moreover, for any SL scheme α, the values of t_0 and T will be taken so that the log-SNR evaluated at times t_0 and t_K has the same value (the initial log-SNR is taken as -4.0 and the last log-SNR is taken as 5.0). This enables fair comparison across different schedules.

D.1. Exploring new integration schemes

By exploiting the relation given in (9) between the denoiser function and the score of the observation process, which is now assumed to be tractable, the SDE (6) is equivalently defined on $[t_0, T]$ by

$$\mathrm{d}Y_t^\alpha = \tfrac{\dot\alpha(t)}{\alpha(t)}\{Y_t^\alpha + \sigma^2 t \nabla \log p_t^\alpha(Y_t^\alpha)\}\mathrm{d}t + \sigma \mathrm{d}B_t \ , \ Y_0^\alpha = 0 \ . \tag{25}$$

Through this formulation, we have fully removed the problem of score estimation, but the issue of discretization error still remains to sample from $(Y_t^\alpha)_{t\in[t_0,T]}$ with (25). Due to the divergence of the coefficient $t \mapsto \dot\alpha(t)/\alpha(t)$ at time 0 (and time T_{gen} in the finite-time setting), we propose to use the stochastic *Exponential Integrator* (EI) scheme (Durmus & Moulines, 2015) in this setting. Consider a time discretization of the interval $[t_0, T]$ defined by an increasing sequence of timesteps $(t_k)_{k=0}^K$ where $t_K = T$ and $K \geq 1$. Then, the EI scheme applied on the SDE (25) amounts to define a sequence of random variables $\{\tilde Y_{t_k}^\alpha\}_{k=0}^K$ obtained by *exactly* integrating the SDE defined for any $k \in \{0, \dots, K-1\}$ by

$$\mathrm{d}\tilde Y_t^\alpha = \tfrac{\dot\alpha(t)}{\alpha(t)}\{\tilde Y_t^\alpha + \sigma^2 t_k \nabla \log p_{t_k}^\alpha(\tilde Y_{t_k}^\alpha)\}\mathrm{d}t + \sigma \mathrm{d}B_t \ , \ t \in [t_k, t_{k+1}] \ .$$

Although the exact integration is not guaranteed for a general schedule α as defined in Section 3.1, our specific design of the denoising schedule $g(t) = t^{-1/2}\alpha(t)$ in the schemes Geom and Geom-∞ provides tractable computations following the result of Lemma 5. We treat these cases separately below.

EI scheme combined with Geom-∞ localization scheme. In this setting, we recall that $g(t) = t^{\alpha_1/2}$. Then, we have

$$\frac{\dot\alpha(t)}{\alpha(t)} = \frac{\alpha_1 + 1}{2t} \ .$$

Note that $\dot\alpha(t)/\alpha(t) \to 0$ as $t \to 0$. Following Lemma 5, the sequence $\{\tilde Y_{t_k}^\alpha\}_{k=0}^K$ obtained by the EI scheme is defined by the recursion

$$\tilde Y_{t_{k+1}}^\alpha = \left(\frac{t_{k+1}}{t_k}\right)^{\frac{\alpha_1+1}{2}} \tilde Y_{t_k}^\alpha + \left\{\left(\frac{t_{k+1}}{t_k}\right)^{\frac{\alpha_1+1}{2}} - 1\right\} \sigma^2 t_k \nabla \log p_{t_k}^\alpha(\tilde Y_{t_k}^\alpha) + \sigma \sqrt{\frac{t_{k+1}}{\alpha_1 t_k^{\alpha_1}}(t_{k+1}^{\alpha_1} - t_k^{\alpha_1})} Z_{k+1} \ .$$

where $(Z_k)_{k=1}^K$ is distributed according to the standard centered Gaussian distribution. In the *standard* case, *i.e.*, $\alpha_1 = 1$, this simplifies as

$$\tilde Y_{t_{k+1}}^\alpha = \tfrac{t_{k+1}}{t_k} \tilde Y_{t_k}^\alpha + \{\tfrac{t_{k+1}}{t_k} - 1\}\sigma^2 t_k \nabla \log p_{t_k}^\alpha(\tilde Y_{t_k}^\alpha) + \sigma \sqrt{\tfrac{t_{k+1}}{t_k}(t_{k+1} - t_k)} Z_{k+1} \ .$$

EI scheme combined with Geom localization scheme. In this second setting, we recall that $g(t) = t^{\alpha_1/2}(1-t)^{-\alpha_2/2}$ ($T_{\text{gen}} = 1$). Then, we have

$$\frac{\dot\alpha(t)}{\alpha(t)} = \frac{\alpha_1+1}{2t} + \frac{\alpha_2}{2(1-t)} \, .$$

Note that $\dot\alpha(t)/\alpha(t) \to 0$ as $t \to 0$ and $t \to 1$. Following Lemma 5, the sequence $\{\tilde Y^\alpha_{t_k}\}_{k=0}^K$ obtained by the EI scheme is defined by the recursion

$$\begin{aligned}
\tilde Y^\alpha_{t_{k+1}} &= \left(\frac{t_{k+1}}{t_k}\right)^{\frac{\alpha_1+1}{2}} \left(\frac{1-t_k}{1-t_{k+1}}\right)^{\frac{\alpha_2}{2}} \tilde Y^\alpha_{t_k} \\
&\quad + \left\{ \left(\frac{t_{k+1}}{t_k}\right)^{\frac{\alpha_1+1}{2}} \left(\frac{1-t_k}{1-t_{k+1}}\right)^{\frac{\alpha_2}{2}} - 1 \right\} \sigma^2 t_k \nabla \log p^\alpha_{t_k}(\tilde Y^\alpha_{t_k}) \\
&\quad + \frac{\sigma}{\sqrt{\alpha_1}} \frac{t_{k+1}^{\frac{\alpha_1+1}{2}}}{(1-t_{k+1})^{\frac{\alpha_2}{2}}} \sqrt{ {}_2\mathrm{F}_1(-\alpha_1, -\alpha_2, 1-\alpha_1, t_k) t_k^{-\alpha_1} - {}_2\mathrm{F}_1(-\alpha_1, -\alpha_2, 1-\alpha_1, t_{k+1}) t_{k+1}^{-\alpha_1} } Z_{k+1} \, ,
\end{aligned}$$

where ${}_2\mathrm{F}_1$ denotes the hypergeometric function, see (Olver, 2010, Chapter 15), and $(Z_k)_{k=1}^K$ is distributed according to the standard centered Gaussian distribution.

When $(\alpha_1, \alpha_2) = (1, 1)$, this simplifies as

$$\begin{aligned}
\tilde Y^\alpha_{t_{k+1}} &= (\tfrac{t_{k+1}}{t_k})(\tfrac{1-t_k}{1-t_{k+1}})^{\frac12} \tilde Y^\alpha_{t_k} \\
&\quad + \{(\tfrac{t_{k+1}}{t_k})(\tfrac{1-t_k}{1-t_{k+1}})^{\frac12} - 1\} \sigma^2 t_k \nabla \log p^\alpha_{t_k}(\tilde Y^\alpha_{t_k}) \\
&\quad + \sigma \sqrt{ \tfrac{t_{k+1}^2}{1-t_{k+1}} \log(\tfrac{t_k}{t_{k+1}}) + \tfrac{t_{k+1}(t_{k+1}-t_k)}{t_k(1-t_{k+1})} } Z_{k+1} \, .
\end{aligned}$$

When $(\alpha_1, \alpha_2) = (2, 1)$, this simplifies as

$$\begin{aligned}
\tilde Y^\alpha_{t_{k+1}} &= (\tfrac{t_{k+1}}{t_k})^{\frac32}(\tfrac{1-t_k}{1-t_{k+1}})^{\frac12} \tilde Y^\alpha_{t_k} \\
&\quad + \{(\tfrac{t_{k+1}}{t_k})^{\frac32}(\tfrac{1-t_k}{1-t_{k+1}})^{\frac12} - 1\} \sigma^2 t_k \nabla \log p^\alpha_{t_k}(\tilde Y^\alpha_{t_k}) \\
&\quad + \sigma t_{k+1} (\tfrac{t_{k+1}/t_k - 1}{1-t_{k+1}})^{\frac12} \sqrt{ \tfrac{t_k + t_{k+1}}{2 t_k t_{k+1}} - 1 } Z_{k+1} \, .
\end{aligned}$$

When $(\alpha_1, \alpha_2) = (1, 2)$, this simplifies as

$$\begin{aligned}
\tilde Y^\alpha_{t_{k+1}} &= (\tfrac{t_{k+1}}{t_k})^{\frac12}(\tfrac{1-t_k}{1-t_{k+1}})^{\frac32} \tilde Y^\alpha_{t_k} \\
&\quad + \{(\tfrac{t_{k+1}}{t_k})^{\frac12}(\tfrac{1-t_k}{1-t_{k+1}})^{\frac32} - 1\} \sigma^2 t_k \nabla \log p^\alpha_{t_k}(\tilde Y^\alpha_{t_k}) \\
&\quad + \tfrac{\sigma}{1-t_{k+1}} \sqrt{ (t_{k+1}-t_k)\{t_{k+1}^2 + \tfrac{t_{k+1}}{t_k}\} + 2 t_{k+1}^2 \log(\tfrac{t_k}{t_{k+1}}) \} } Z_{k+1} \, .
\end{aligned}$$

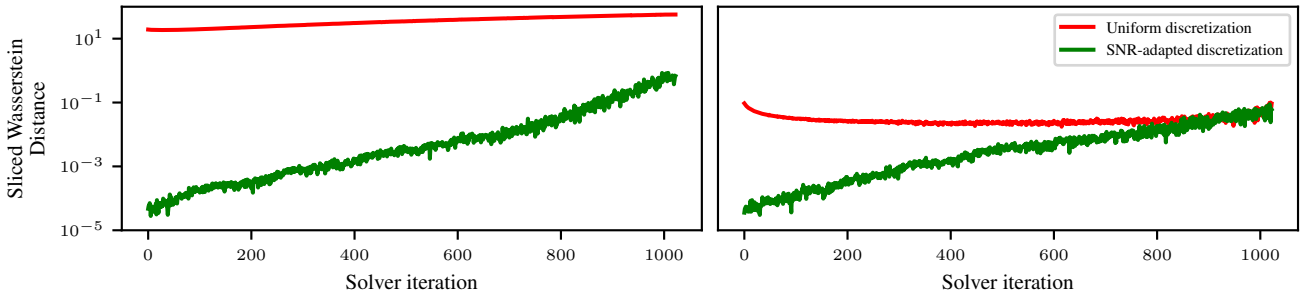

Figure 6: Sliced Wasserstein distance between the true samples from $p^{\alpha}_{t_k}$ and the empirical distribution of $\tilde{Y}^{\alpha}_{t_k}$ obtained by EI scheme combined with uniform time discretization (green) and SNR-adapted discretization (red) for the Standard (*left*) and Geom(1,1) (*right*) schemes.

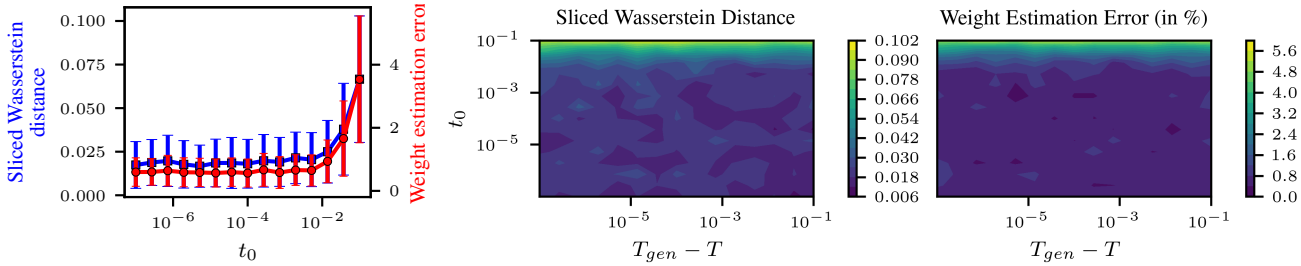

Figure 7: Impact of t_0 with different schedules measured with the relative weight estimation error and the sliced Wasserstein distance. **Left**: Impact of t_0 in the standard scheme. **Right**: Impact of t_0 and $(T_{\text{gen}} - T)$ in the Geom(1,1) scheme.

D.2. Studying the impact of the SNR-adapted time discretization

In Section 4.1, we suggested to take a SNR-adapted time discretization by $(t_k)_{k=0}^K$ such that

$$\log \mathrm{SNR}(t_k) = \log \mathrm{SNR}(t_0) + \Delta_{\mathrm{SNR}} k \ ,$$

with $t_0 > 0$ and $\Delta_{\mathrm{SNR}} = (\log \mathrm{SNR}(T) - \log \mathrm{SNR}(t_0))/K$. Figure 6 shows that the SNR-adapted discretization efficiently reduces the integration error for both schedules Geom and Geom-∞. Note, the impact on the schedule Geom(1,1) is more moderated as the uniform initialization already splits the curve into moderated log-SNR increments (see Figure 2) because of a slower rate near t_0.

D.3. Studying the impact of the Gaussian approximation in the initialization

In this section only, we replace the target distribution considered in (24) by the mixture of two Gaussian distributions $\mathsf{N}(-3\,\mathbf{1}_{10}, \Sigma)$ and $\mathsf{N}(3\,\mathbf{1}_{10}, \Sigma)$, where $\Sigma = 0.05\,\mathrm{I}_{10}$, with weights respectively given by $2/3$ and $1/3$. Therefore, π has non-zero mean and high variance, which provides a challenging setting for our Gaussian approximation at initialization given by $Y^{\alpha}_{t_0} \sim \mathsf{N}(0, \sigma^2 t_0\,\mathrm{I}_d)$. Without information on the target, this is the best estimation that we have when t_0 is close to 0. Figure 7 shows that only high values of t_0 (*i.e.* above 10^{-2}) degrades the performance in our localization schemes. This underlines the need of running the Langevin-within-Langevin correction procedure of this approximation as explained in Section 4.3. Additionally, note that Figure 7 (*right*) shows that taking $(T_{\text{gen}} - T)$ low in the finite time setting does not degrade performance, which was not obvious from the log-SNR shape near T_{gen} in Figure 1.

D.4. Studying the impact of the number of discretization steps

The bottom row of Figure 8 shows that, in the ideal case where the score is known, using the SNR-adapted discretization makes all the different schemes equivalent[8]. Moreover, we see that the sampling error quickly stabilizes with medium budgets which highlights that the integration error was successfully minimized.

[8]We recall that we choose t_0 and T so that the starting and ending levels of SNR are the same across different schemes.

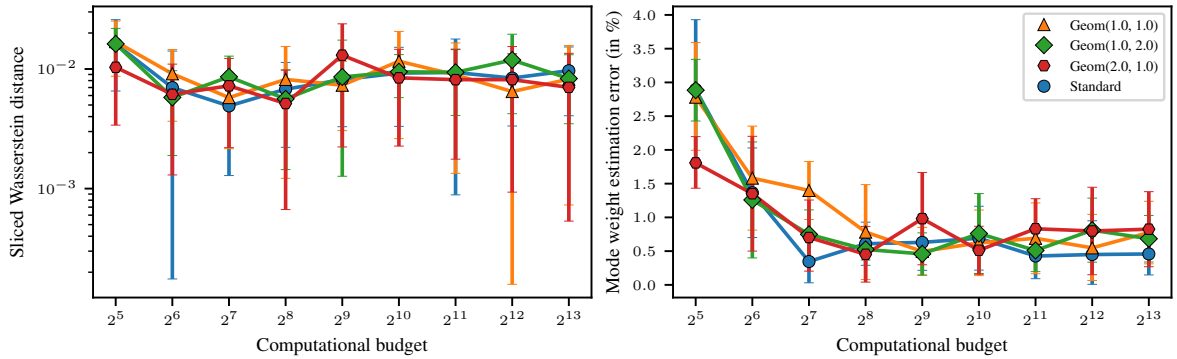

Figure 8: Impact of the computational budget K with different schemes - The sampling error is computed either with the Sliced Wasserstein distance (*left*) or the relative weight error (*right*).

E. Theoretical results on the duality of log-concavity

In this section, we detail the results on log-concavity provided in Section 4.2 and Section 4.3. For sake of readability, those are stated with a general $\sigma > 0$. Our first result provides uniform upper bounds on the Hessian of the log-density of the observation process and its corresponding log-posterior density.

Lemma 15. *Assume A0. Let $t \in (0, T_{gen})$. We recall that p_t^α stands for the marginal distribution at time t of the observation process defined by (4), while q_t^α stands for the corresponding posterior density. Under regularity assumptions on π detailed in the proof of the lemma, we have for any $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ that*

$$\nabla_y^2 \log p_t^\alpha(y) \preccurlyeq \zeta_p(t)\, \mathrm{I}_d \;, \quad \textit{where } \zeta_p(t) = \tfrac{\alpha(t)^2 dR^2}{(\alpha(t)^2\tau^2+\sigma^2 t)^2} - \tfrac{1}{\alpha(t)^2\tau^2+\sigma^2 t} \;, \tag{26}$$

$$\nabla_x^2 \log q_t^\alpha(x|y) \preccurlyeq \zeta_q(t)\, \mathrm{I}_d \;, \quad \textit{where } \zeta_q(t) = \tfrac{dR^2}{\tau^4} - \tfrac{1}{\tau^2} - \tfrac{g(t)^2}{\sigma^2} \;. \tag{27}$$

In particular, ζ_q is a strictly decreasing function on $(0, T_{gen})$.

Proof. We begin by proving (26). Consider the stochastic process $(Y_t^\alpha)_{t\in[0,T_{\mathrm{gen}})}$, given in (4), and defined by $Y_t^\alpha = \alpha(t)X + \sigma W_t$ where $X \sim \pi$ and $(W_t)_{t\geq 0}$ is a standard Brownian motion. Due to **A0**, X can be written as $X = U + G$ where $U \sim \mu$, with μ compactly supported on $\mathsf{B} = \mathrm{B}(\mathbf{m}_\pi, R\sqrt{d})$, and $G \sim \mathsf{N}(0, \tau^2\, \mathrm{I}_d)$. In particular, it holds almost surely that $\|U - \mathbf{m}_\pi\|^2 \leq dR^2$. We thus have the following decomposition

$$p_t^\alpha(y) = \textstyle\int_{\mathsf{B}} p_t^\alpha(y|u)\mathrm{d}\mu(u) \;,$$

where $p_t^\alpha(\cdot|u)$ is the conditional density of Y_t^α given $U = u \in \mathsf{B}$ defined as

$$p_t^\alpha(y|u) = \mathsf{N}(y; \alpha(t)u, \{\alpha(t)^2\tau^2 + \sigma^2 t\}\, \mathrm{I}_d) \;. \tag{28}$$

Then, it comes that

$$\nabla_y \log p_t^\alpha(y|u) = \frac{-y + \alpha(t)u}{\alpha(t)^2\tau^2 + \sigma^2 t} \;, \; \nabla_y^2 \log p_t^\alpha(y|u) = -\frac{1}{\alpha(t)^2\tau^2 + \sigma^2 t}\mathrm{I}_d \;.$$

We now make the technical assumption that for any $k \in \{1, 2\}$, there exists $\varphi_{p,k} : \mathsf{B} \to \mathbb{R}_+$ such that $\int_{\mathsf{B}} \varphi_{p,k}(u)\mathrm{d}\mu(u) < \infty$ and for any $(u, y) \in \mathsf{B} \times \mathbb{R}^d$, we have $\left\|\nabla_y^k p_t^\alpha(y|u)\right\| \leq \varphi_{p,k}(u)$. By combining this assumption with the result of Lemma 4-(18), we obtain that

$$\begin{aligned}\nabla_y^2 \log p_t^\alpha(Y_t^\alpha) &= -\frac{1}{\alpha(t)^2\tau^2 + \sigma^2 t}\mathrm{I}_d + \mathrm{Cov}\left[\frac{-Y_t^\alpha + \alpha(t)U}{\alpha(t)^2\tau^2 + \sigma^2 t}\middle| Y_t^\alpha\right] \\ &= -\frac{1}{\alpha(t)^2\tau^2 + \sigma^2 t}\mathrm{I}_d + \frac{\alpha(t)^2}{(\alpha(t)^2\tau^2 + \sigma^2 t)^2}\mathrm{Cov}[U|Y_t^\alpha] \;.\end{aligned}$$

Since $\|U - \mathbf{m}_\pi\|^2 \leq dR^2$, we have $\mathrm{Cov}[U|Y_t^\alpha] \preccurlyeq dR^2\,\mathrm{I}_d$[9]. Then, we obtain the bound (26) as $\nabla^2 \log p_t^\alpha$ is continuous and p_t^α is positive on $\mathbb{R}^d$.

We now prove (27). We recall from (8) that $q_t^\alpha(x|y) \propto \pi(x)\mathsf{N}(x; y/\alpha(t), \sigma^2/g(t)^2\,\mathrm{I}_d)$. Therefore

$$\nabla_x^2 \log q_t^\alpha(X|Y_t^\alpha) = \nabla_x^2 \log \pi(X) - \frac{g(t)^2}{\sigma^2}\mathrm{I}_d \;.$$

In particular, we have $\pi(x) = \int_{\mathsf{B}} \mathsf{N}(x; u, \tau^2\,\mathrm{I}_d)\mathrm{d}\mu(u)$.

We now make the technical assumption that for any $k \in \{1, 2\}$, there exists $\varphi_{q,k} : \mathsf{B} \to \mathbb{R}_+$ such that $\int_{\mathsf{B}} \varphi_{q,k}(u)\mathrm{d}\mu(u) < \infty$ and for any $(u, x) \in \mathsf{B} \times \mathbb{R}^d$, we have $\left\|\nabla_x^k \mathsf{N}(x; u, \tau^2\,\mathrm{I}_d)\right\| \leq \varphi_{p,k}(u)$. Then, by using again the result of Lemma 4-(18), we obtain that

$$\nabla_x^2 \log \pi(X) = -\frac{1}{\tau^2}\mathrm{I}_d + \frac{1}{\tau^4}\mathrm{Cov}[U|X] \;,$$

Since $\|U - \mathbf{m}_\pi\|^2 \leq dR^2$, we also have $\mathrm{Cov}[U|X] \preccurlyeq dR^2\,\mathrm{I}_d$[10]. We obtain (27) with similar reasoning as before. □

Note that the upper bounds obtained in Lemma 15 can be made tighter in the case where the target distribution is a Gaussian mixture, as shown in Lemma 16.

Lemma 16. *Let $a > 0$, $\gamma > 0$ and $w \in (0, 1)$. Consider the target distribution π defined as the mixture of the Gaussian distributions $\mathsf{N}(-a\,\mathbf{1}_d, \gamma^2\,\mathrm{I}_d)$ and $\mathsf{N}(+a\,\mathbf{1}_d, \gamma^2\,\mathrm{I}_d)$, with respective weights w and $1 - w$. Then for any $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, we have*

$$\nabla_y^2 \log p_t^\alpha(y) = \tfrac{2\alpha(t)^2 a^2}{(\alpha(t)^2\gamma^2 + \sigma^2 t)^2(1 + \cosh g(t,y))}\,\mathbf{1}_d\mathbf{1}_d^\top - \tfrac{1}{\alpha(t)^2\gamma^2 + \sigma^2 t}\,\mathrm{I}_d \;, \text{ with } g(t, y) = \tfrac{2\alpha(t) a y^\top \mathbf{1}_d}{\alpha(t)^2\gamma^2 + \sigma^2 t} + \log(\tfrac{1}{w} - 1) \quad (29)$$

$$\nabla_x^2 \log q_t^\alpha(x|y) = \tfrac{2a^2}{\gamma^4(1 + \cosh f(t,x))}\,\mathbf{1}_d\mathbf{1}_d^\top - \tfrac{1}{\gamma^2}\,\mathrm{I}_d - \tfrac{g(t)^2}{\sigma^2}\,\mathrm{I}_d \;, \text{ with } f(t, x) = \tfrac{2a x^\top \mathbf{1}_d}{\gamma^2} + \log(\tfrac{1}{w} - 1) \quad (30)$$

In particular, tighter expressions of ζ_p and ζ_q can be obtained in Lemma 15 by replacing R by $R/\{2\max(w, 1 - w)\} \leq R$.

The proof of this result is inspired from the derivation of (Saremi et al., 2024, Eq. (4.6)).

Proof. As explained in Appendix C, π verifies **A**0, where $\tau = \gamma$ and μ is a mixture of two Dirac masses at $-a\,\mathbf{1}_d$ and $a\,\mathbf{1}_d$ with respective weights w and $1 - w$, whose density is given by $p(u) = w\mathbb{1}_{-a\,\mathbf{1}_d}(u) + (1 - w)\mathbb{1}_{a\,\mathbf{1}_d}(u)$.

We recall that the stochastic process $(Y_t^\alpha)_{t\in[0,T_{\text{gen}})}$, given in (4), is defined by $Y_t^\alpha = \alpha(t)X + \sigma W_t$ where $X \sim \pi$ and $(W_t)_{t\geq 0}$ is a standard Brownian motion. Due to **A**0, X can be written as $X = U + G$ where $U \sim \mu$ and $G \sim \mathsf{N}(0, \gamma^2\,\mathrm{I}_d)$.

We begin by proving (29). Following the first part of the proof of Lemma 15, we have

$$\nabla_y^2 \log p_t^\alpha(y) = -\frac{1}{\alpha(t)^2\gamma^2 + \sigma^2 t}\mathrm{I}_d + \frac{\alpha(t)^2}{(\alpha(t)^2\gamma^2 + \sigma^2 t)^2}\mathrm{Cov}[U_y] \;,$$

where U_y is a random vector distributed according to $p_t^\alpha(\cdot|y)$, the conditional distribution of U given $Y_t^\alpha = y \in \mathbb{R}^d$, whose density is given by

$$p_t^\alpha(u|y) \propto p(u)p_t^\alpha(y|u) \;,$$

where $p_t^\alpha(y|u)$ is defined in (28). Therefore, we obtain that

$$p_t^\alpha(u|y) \propto \underbrace{w \exp\left(-\frac{\|y + \alpha(t)a\mathbf{1}_d\|^2}{2\{\alpha(t)^2\gamma^2 + \sigma^2 t\}}\right)}_{w_1(t,y)}\mathbf{1}_{-a\,\mathbf{1}_d}(u) + \underbrace{(1 - w)\exp\left(-\frac{\|y - \alpha(t)a\mathbf{1}_d\|^2}{2\{\alpha(t)^2\gamma^2 + \sigma^2 t\}}\right)}_{w_2(t,y)}\mathbf{1}_{a\,\mathbf{1}_d}(u) \;.$$

[9] Note that this upper bound may be loose, since it does not depend on y.
[10] Note that this upper bound may be loose, since it does not depend on x.

Hence, $p_t^\alpha(\cdot|y)$ is a mixture of two Dirac masses at $-a\,\mathbf{1}_d$ and $a\,\mathbf{1}_d$ with respective weights $\tilde{w}_1(t,y) = w_1(t,y)/\{w_1(t,y) + w_2(t,y)\}$ and $\tilde{w}_2(t,y) = w_2(t,y)/\{w_1(t,y) + w_2(t,y)\}$. Therefore, we get that

$$\begin{aligned}\mathrm{Cov}[U_y] &= \mathbb{E}[U_y U_y^\top] - \mathbb{E}[U_y]\mathbb{E}[U_y]^\top \\ &= \tilde{w}_1(t,y) a^2 \mathbf{1}_d \mathbf{1}_d^\top + \tilde{w}_2(t,y) a^2 \mathbf{1}_d \mathbf{1}_d^\top - (\tilde{w}_2(t,y) - \tilde{w}_1(t,y))^2 a^2 \mathbf{1}_d \mathbf{1}_d^\top \\ &= \{1 - (\tilde{w}_2(t,y) - \tilde{w}_1(t,y))^2\} a^2 \mathbf{1}_d \mathbf{1}_d^\top \,.\end{aligned}$$

In particular, we have

$$\tilde{w}_2(t,y) - \tilde{w}_1(t,y) = \frac{w_2(t,y) - w_1(t,y)}{w_2(t,y) + w_1(t,y)} = \frac{\exp(\log w_2(t,y)) - \exp(\log w_1(t,y))}{\exp(\log w_2(t,y)) - \exp(\log w_1(t,y))} \,,$$

where $\log w_1(t,y) = -\frac{\|y + \alpha(t) a \mathbf{1}_d\|^2}{2\{\alpha(t)^2\gamma^2 + \sigma^2 t\}} + \log(w)$ and $\log w_2(t,y) = -\frac{\|y - \alpha(t) a \mathbf{1}_d\|^2}{2\{\alpha(t)^2\gamma^2 + \sigma^2 t\}} + \log(1-w)$. Then, it comes that

$$\begin{aligned}1 - (\tilde{w}_2(t,y) - \tilde{w}_1(t,y))^2 &= 1 - \tanh\left(\frac{\log w_2(t,y) - \log w_1(t,y)}{2}\right) \\ &= \frac{2}{1 + \cosh(\log w_2(t,y) - \log w_1(t,y))} \,.\end{aligned}$$

Moreover, we get that

$$\log w_2(t,y) - \log w_1(t,y) = \frac{2\alpha(t) a y^\top \mathbf{1}_d}{\alpha(t)^2\gamma^2 + \sigma^2 t} + \log\left(\frac{1}{w} - 1\right) \,.$$

Denote $g(t,y) = \log w_2(t,y) - \log w_1(t,y)$. We finally obtain (29) by combining previous computations.

We now prove (30). Following the second part of the proof of Lemma 15, we have

$$\nabla_x^2 \log q_t^\alpha(x|y) = -\frac{1}{\gamma^2}\mathrm{I}_d - \frac{g(t)^2}{\gamma^2}\mathrm{I}_d + \frac{1}{\gamma^4}\mathrm{Cov}[U_x] \,,$$

where U_x is a random vector distributed according to $\mu(\cdot|x)$, the conditional distribution of U given $X = x \in \mathbb{R}^d$, whose density is given by

$$\mu(u|x) \propto \mathsf{N}(x; u, \gamma^2\,\mathrm{I}_d)\mu(u) \,.$$

Therefore, we obtain that

$$\mu(u|x) \propto \underbrace{w \exp\left(-\frac{\|x + a\mathbf{1}_d\|^2}{2\gamma^2}\right)}_{w_1(t,x)} \mathbf{1}_{-a\,\mathbf{1}_d}(u) + \underbrace{(1-w)\exp\left(-\frac{\|x - a\mathbf{1}_d\|^2}{2\gamma^2}\right)}_{w_2(t,x)} \mathbf{1}_{a\,\mathbf{1}_d}(u) \,.$$

Hence, $\mu(\cdot|x)$ is a mixture of two Dirac masses at $-a\,\mathbf{1}_d$ and $a\,\mathbf{1}_d$ with respective weights $\tilde{w}_1(t,x) = w_1(t,x)/\{w_1(t,x) + w_2(t,x)\}$ and $\tilde{w}_2(t,x) = w_2(t,x)/\{w_1(t,x) + w_2(t,x)\}$. Therefore, we get that

$$\begin{aligned}\mathrm{Cov}[U_x] &= \mathbb{E}[U_x U_x^\top] - \mathbb{E}[U_x]\mathbb{E}[U_x]^\top \\ &= \tilde{w}_1(t,x) a^2 \mathbf{1}_d \mathbf{1}_d^\top + \tilde{w}_2(t,x) a^2 \mathbf{1}_d \mathbf{1}_d^\top - (\tilde{w}_2(t,x) - \tilde{w}_1(t,x))^2 a^2 \mathbf{1}_d \mathbf{1}_d^\top \\ &= \{1 - (\tilde{w}_2(t,x) - \tilde{w}_1(t,x))^2\} a^2 \mathbf{1}_d \mathbf{1}_d^\top \,.\end{aligned}$$

Similarly to the computations derived above, we obtain that

$$1 - (\tilde{w}_2(t,x) - \tilde{w}_1(t,x))^2 = \frac{2}{1 + \cosh f(t,x)} \,,$$

where

$$f(t,x) = \log w_2(t,x) - \log w_1(t,x) = \frac{2a x^\top \mathbf{1}_d}{\gamma^2} + \log\left(\frac{1}{w} - 1\right) \,,$$

which finally leads to (30).

We now derive tighter expressions for ζ_p and ζ_q in this setting. Since $2/(1+\cosh) \leq 1$ and $\mathbf{1}_d \mathbf{1}_d^\top \preccurlyeq d\,\mathrm{I}_d$, we have

$$\nabla_y^2 \log p_t^\alpha(y) \preccurlyeq \tfrac{\alpha(t)^2 a^2 d}{(\alpha(t)^2\gamma^2+\sigma^2 t)^2}\,\mathrm{I}_d - \tfrac{1}{\alpha(t)^2\gamma^2+\sigma^2 t}\,\mathrm{I}_d \;,$$
$$\nabla_x^2 \log q_t^\alpha(x|y) \preccurlyeq \tfrac{a^2 d}{\gamma^4}\,\mathrm{I}_d - \tfrac{1}{\gamma^2}\,\mathrm{I}_d - \tfrac{g(t)^2}{\sigma^2}\,\mathrm{I}_d \;.$$

In this particular setting, we recall that $R = 2\max(w, 1-w)a$, see Appendix C, and thus, the upper bounds on the Hessians derived above are equivalent to

$$\nabla_y^2 \log p_t^\alpha(y) \preccurlyeq \tfrac{\alpha(t)^2 d\tilde{R}^2}{(\alpha(t)^2\gamma^2+\sigma^2 t)^2}\,\mathrm{I}_d - \tfrac{1}{\alpha(t)^2\gamma^2+\sigma^2 t}\,\mathrm{I}_d \;,$$
$$\nabla_x^2 \log q_t^\alpha(x|y) \preccurlyeq \tfrac{d\tilde{R}^2}{\gamma^4}\,\mathrm{I}_d - \tfrac{1}{\gamma^2}\,\mathrm{I}_d - \tfrac{g(t)^2}{\sigma^2}\,\mathrm{I}_d \;,$$

where $\tilde{R} = R/\{2\max(w, 1-w)\} \leq R$. Therefore, we can obtain tighter expressions of ζ_p and ζ_q in Lemma 15 by replacing R by $\tilde{R}$. □

We provide below a result, that combines formal versions of Theorem 1 and Theorem 2.

Proposition 17. *Assume A0. Denote by g^{-1} the inverse function of the denoising schedule g.*

If $dR^2 > \tau^2$, define

$$t_{\mathrm{p}} = g^{-1}\left(\tfrac{\sigma}{\sqrt{dR^2-\tau^2}}\right) \quad \textit{and} \quad t_{\mathrm{q}} = g^{-1}\left(\tfrac{\sigma\sqrt{dR^2-\tau^2}}{\tau^2}\right) \;,$$

otherwise, define $t_{\mathrm{p}} = T_{gen}$ and $t_{\mathrm{q}} = 0$. Then,

(a) for any $y \in \mathbb{R}^d$, $q_t^\alpha(\cdot|y)$ is strongly log-concave for $t \in (t_{\mathrm{q}}, T_{gen})$ and gets more log-concave as t increases,

(b) p_t^α is strongly log-concave for $t \in (0, t_{\mathrm{p}})$.

Proof. Consider the expressions of t_{p} and t_{q} given above. Note that they are well defined in the case where $dR^2 > \tau^2$ since g is a bijection from $[0, T_{\mathrm{gen}})$ to $\mathbb{R}_+$. We begin with the proof of the result (a). Let $t \in (0, T_{\mathrm{gen}})$. Following Lemma 15-(27), for any $y \in \mathbb{R}^d$, $q_t^\alpha(\cdot|y)$ is strongly log concave if

$$\zeta_q(t) < 0 \iff g(t)^2 > \frac{\sigma^2(dR^2-\tau^2)}{\tau^4} \iff t > t_{\mathrm{q}} \;.$$

Moreover, q_t^α gets more log-concave as t increases as ζ_q is a strictly decreasing function, see Lemma 15. We now give the proof of the result (b). Let $t \in (0, T_{\mathrm{gen}})$. Following Lemma 15-(26), p_t^α is strongly log concave if

$$\zeta_p(t) < 0 \iff \frac{\alpha(t)^2 dR^2}{\alpha(t)^2\tau^2+\sigma^2 t} < 1 \iff g(t)^2(dR^2-\tau^2) < \sigma^2 \iff t < t_{\mathrm{p}} \;.$$

□

Hence, this result shows that the condition $dR^2 > \tau^2$ is restrictive on the log-concavity of the marginal distribution and the posterior of the localization model. In terms of Gaussian mixtures, this condition can be interpreted as having the distance between the modes that is larger than the variance of the modes. We now restate Theorem 3 and give its proof.

Proposition 18. *Assume A0, where $dR^2 < 2\tau^2$. Then, $t_{\mathrm{q}} < t_{\mathrm{p}}$, where t_{q} and t_{p} are defined in Proposition 17.*

Proof. If $dR^2 \leq \tau^2$, this result is directly obtained since $t_{\mathrm{q}} = 0$ and $t_{\mathrm{p}} = T_{\mathrm{gen}}$. Assume that $dR^2 > \tau^2$. We have

$$t_{\mathrm{q}} < t_{\mathrm{p}} \iff \frac{dR^2-\tau^2}{\tau^4} < \frac{1}{dR^2-\tau^2} \iff (dR^2-\tau^2)^2 < \tau^4 \iff dR^2 < 2\tau^2 \;,$$

which gives the result. □

Hence, Proposition 18 shows that a sweet spot for the hyper-parameter t_0 in SLIPS exists under a restrictive condition on R and τ in Assumption **A**0, enabling the duality of log-concavity. Elaborating on the result of Lemma 16, this condition may be slightly alleviated in the case of an *unbalanced* Gaussian mixture as proved in Proposition 19.

Proposition 19. *Let $a > 0$, $\gamma > 0$ and $w \in (0,1)$. Consider the target distribution π defined as the mixture of the Gaussian distributions $\mathsf{N}(-a\,\mathbf{1}_d, \gamma^2\,\mathrm{I}_d)$ and $\mathsf{N}(+a\,\mathbf{1}_d, \gamma^2\,\mathrm{I}_d)$, with respective weights w and $1-w$. Assume that $dR^2 < 2\tau^2 \times 4\max(w, 1-w)^2$. Then, the duality of log-concavity is ensured to exist in SLIPS.*

Proof. This result can be seen as an extension of Proposition 17 and Proposition 18, in the case where R is replaced by $R/\{2\max(w, 1-w)\}$, see Lemma 16. □

F. Details on SLIPS algorithm

F.1. Details on the implementation of SLIPS

Langevin-within-Langevin initialization. Consider a timestep t_0 well chosen such that $p^\alpha_{t_0}$ and $q^\alpha_{t_0}$ are both approximately log-concave. The goal of our initialization is to sample from $p^\alpha_{t_0}$ at lowest cost. Thanks to log-concavity of $p^\alpha_{t_0}$, we may consider to apply ULA to sample from this distribution. Given $N \geq 1$ and a step-size $\lambda > 0$, this amounts to consider a sequence of random variables $\{Y^{(n)}\}_{n=0}^N$, defined by the following recursion

$$Y^{(n+1)} = Y^{(n)} + \lambda \nabla \log p^\alpha_{t_0}(Y^{(n)}) + \sqrt{2\lambda} Z^{(n+1)} \,, \tag{31}$$

where $\{Z^{(n)}\}_{n=1}^N$ are independently distributed according to the standard centered Gaussian distribution. We recall that the score $\nabla \log p^\alpha_{t_0}$ is however not tractable but can be re-expressed with (9) such that for any $y \in \mathbb{R}^d$, we have

$$\nabla \log p^\alpha_{t_0}(y) = \{\alpha(t_0) u^\alpha_{t_0}(y) - y\}/\sigma^2 t_0 \,,$$

where $u^\alpha_{t_0}(y)$ is the expectation of the posterior $q^\alpha_{t_0}(\cdot|y)$ given in (8). Although this term is intractable too, it can be estimated by approximately sampling from $q^\alpha_{t_0}$ via MALA, since $q^\alpha_{t_0}$ is also ensured to be approximately log-concave.

Building upon this relation, at each step $n \in \{0, \dots, N-1\}$, we propose to approximate $\nabla \log p^\alpha_{t_0}(Y^{(n)})$ in the recursion (31) with a MCMC-based estimator of $u^\alpha_{t_0}(Y^{(n)})$ obtained by sampling from $q^\alpha_{t_0}(\cdot|Y^{(n)})$. This results in the Langevin-within-Langevin procedure summarized in Algorithm 2. We finally highlight three main choices in the design of this algorithm: (i) we choose the Gaussian approximation $Y^{(0)} \sim \mathsf{N}(0, \sigma^2 t_0\,\mathrm{I}_d)$ for ULA initialization, relying on the study conducted in Appendix D.3; (ii) consequently, we set the step-size λ to be slightly smaller than the variance of this distribution, *i.e.*, $\lambda = \sigma^2 t_0/2$; (iii) the MALA-based posterior sampling is initialized with the best estimation of $u^\alpha_{t_0}(Y^{(n)})$ that we have, *i.e.*, $Y^{(n)}/\alpha(t_0)$, recalling that $Y^\alpha_t/\alpha(t)$ and $u^\alpha_t(Y^\alpha_t)$ have the same localization behaviour, see Appendix B.2.

Algorithmic technicalities. We now present three algorithmic subtleties featured in SLIPS.

(a) The step-size of MALA used to sample from the posterior is adapted. Using the acceptance ratio of the Metropolis-Hasting filter, we geometrically decrease (respectively increase) the step-size when the acceptance ratio is below (respectively above) a target ratio of 75%;

(b) The last MCMC samples obtained when sampling the posterior at step k will be the first samples when sampling at step $k+1$. This persistent trick is motivated by the fact that the posterior is expected to change little from one iteration to the next. This behaviour is notably illustrated in Figure 3 *(bottom row)*.

(c) Lastly, as mentioned and justified in Appendix B.2, we use the estimated denoiser $\tilde{U}^\alpha_T$ at the final integration timestep T as an approximate sample from π rather than $Y^\alpha_T/\alpha(T)$ to align with related work.

F.2. Ablation study

In this section, we investigate the impact of the different hyper-parameters of SLIPS. Similarly to Appendix D, we run our study by applying our algorithm on the unbalanced bimodal Gaussian mixture defined in (24) with $d = 10$. This section will clarify three points:

(a) The behaviour of SLIPS within the assumptions of Theorem 3;

(b) The existence of a "sweet spot" for t_0 outside the assumptions of Theorem 3 and the impact of the estimation of R_π;

(c) The tuning of hyper-parameters in the Langevin-within-Langevin initialization (see Algorithm 2).

Behaviour of `SLIPS` within the restrictive assumptions. In this part only, we consider the case of a mixture of two Gaussian distributions in dimension 5 given by $\mathsf{N}(-(2/3)a\,\mathbf{1}_5, \tau^2 \mathrm{I}_5)$ and $\mathsf{N}(+(4/3)a\,\mathbf{1}_5, \tau^2 \mathrm{I}_5)$, where $\tau^2 = 0.1$ and $a > 0$ is *not fixed*, with respective weights $2/3$ and $1/3$. By varying a, we vary R, recalling that $R = (4/3)a$, see Appendix C. In Figure 9, we run `SLIPS` with different schemes while setting t_0 as the mean value between t_p and t_q derived in Proposition 17, in accordance with Theorem 3 (*left*), or derived by combining the results of Proposition 17 and Proposition 19, which is less restrictive (*right*). When p_t^α and q_t^α are both log-concave independently of t, *i.e.*, when $dR^2/\tau^2 \leq 1$ for the general case (see Proposition 17) or $dR^2/\tau^2 \leq 16/9$ in the refined Gaussian mixture setting (see Lemma 16), we always set $t_0 = 10^{-2}$. This choice ensures that the Gaussian approximation which corresponds to the initialization of the Langevin-within-Langevin algorithm is accurate (see Figure 7 for details). This figure shows that choosing t_0 according to Proposition 17 leads to accurate sampling below the threshold and degraded (or equal) performance above this threshold.

Existence of a sweet spot for duality outside of the restrictive assumptions. Given the target distribution described in (24) with $d = 10$, we have that $R = 4/3$ and $\tau^2 = 0.05$, which means that $R^2 d/\tau^2 \approx 350$. Hence, the distribution at stake does not fit the additional assumption of Theorem 3, or even the refined assumption of Proposition 19. However, Figure 10 shows that there still exists a sweet spot for t_0 in this case. Moreover, we observe in Figure 11 that a poor estimation of R_π shifts the sweet spot. This makes sense as the sweet spot likely corresponds to a SNR level. Recalling that $\log \mathrm{SNR}(t) = 2\log(g(t)) + \log(R_\pi^2/(\sigma^2 d))$, if $\sigma > R_\pi/\sqrt{d}$ (*i.e.*, we overestimate R_π) the log-SNR is shifted downwards and the resulting t_0 should be larger (as we see in the right column of Figure 11); if $\sigma < R_\pi/\sqrt{d}$ (*i.e.*, we underestimate R_π), the resulting t_0 should be smaller (as we see in the left column of Figure 11).

Choice of hyper-parameters in the Langevin-within-Langevin initialization. Figure 12 shows that only a few steps in the Langevin-within-Langevin algorithm (see Algorithm 2) are needed to reach a stationary sampling error.

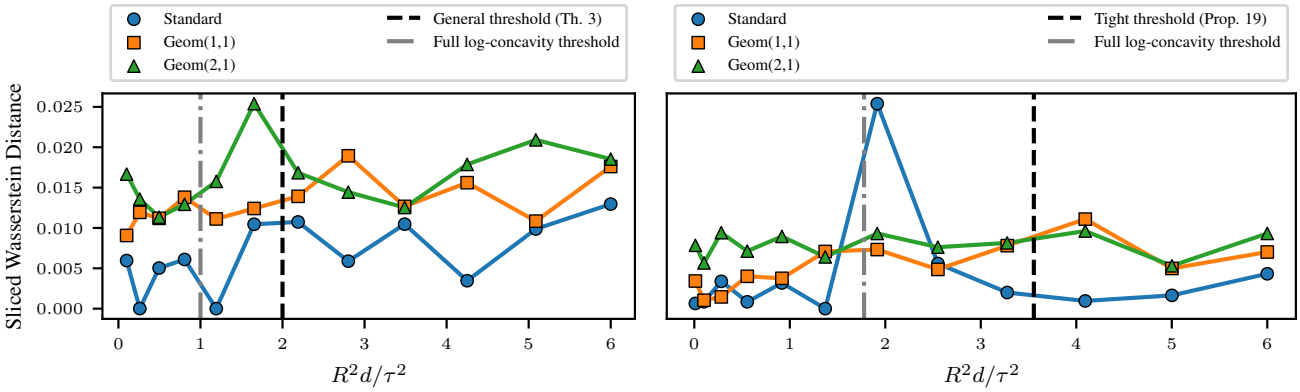

Figure 9: Sliced Wasserstein distance depending on the ratio $R^2 d/\tau^2$. **Left**: $R^2 d/\tau^2 = 2$ is the threshold of duality of log-concavity in Theorem 3 (*general case*). **Right**: $R^2 d/\tau^2 = 32/9$ is the threshold of duality of log-concavity in Proposition 19 (*refinement*). In both cases, the grey line corresponds to the threshold where the log-concavity conditions start to be restrictive on the marginal distribution of the observation process and the corresponding posterior distribution.

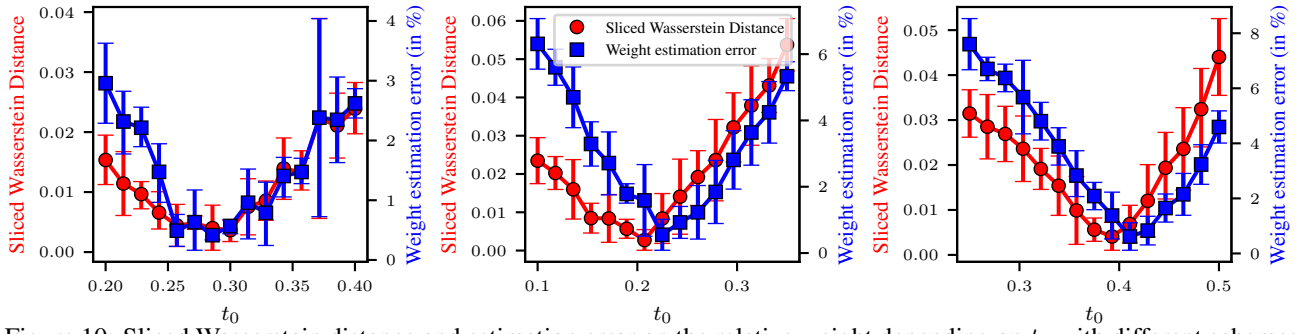

Figure 10: Sliced Wasserstein distance and estimation error on the relative weight depending on t_0 with different schemes. **Left**: Standard. **Middle**: Geom(1,1). **Right**: Geom(2,1).

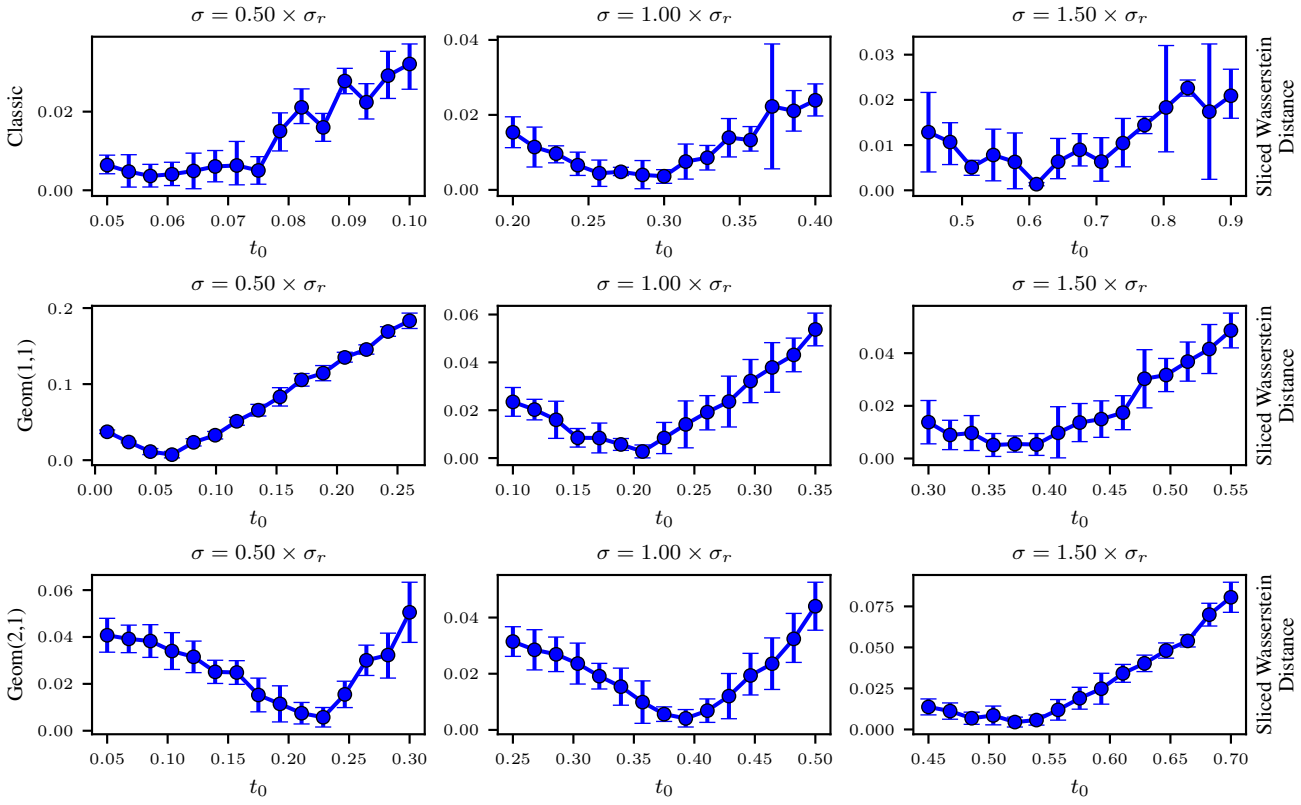

Figure 11: Sliced Wasserstein distance depending on t_0 for different values of σ. We denote $\sigma_r = R_\pi/\sqrt{d}$.

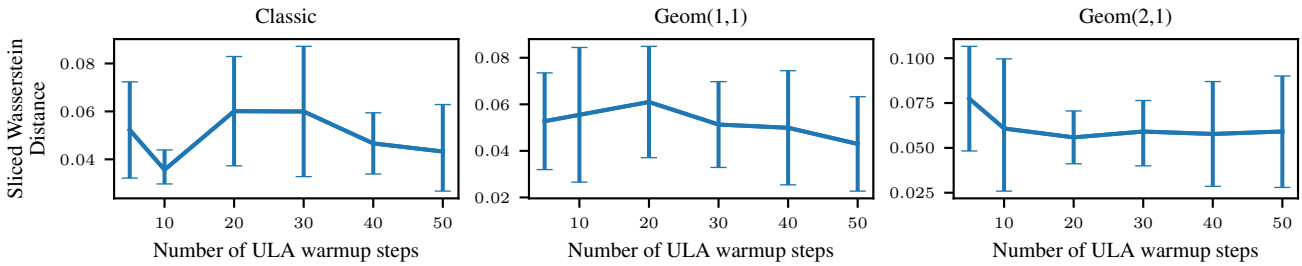

Figure 12: Sliced Wasserstein distance depending on the number of Langevin-within-Langevin steps.

G. Details on related work

In this appendix, we provide further details on related works, highlighting similarities and identifying the main limitations.

G.1. Reverse Diffusion Monte Carlo (Huang et al., 2024)

Given $T > 0$, the authors consider the denoising process $(Y_t)_{t\in[0,T]}$ that is solution to the following SDE

$$\mathrm{d}Y_t = \{Y_t + 2\nabla \log p_{T-t}(Y_t)\} + \sqrt{2}\mathrm{d}B_t \ , \ Y_0 \sim p_T \ , \tag{32}$$

where $(B_t)_{t\geq 0}$ is a Brownian motion in $\mathbb{R}^d$ and p_s is the *intractable* marginal distribution defined for any $s > 0$ and any $y \in \mathbb{R}^d$ by $p_s(y) = \int_{\mathbb{R}^d} \mathsf{N}(y; \mathrm{e}^{-s}x, (1-\mathrm{e}^{-2s})\,\mathrm{I}_d)\mathrm{d}\pi(x)$. Under mild conditions on π (Cattiaux et al., 2023), $(Y_t)_{t\in[0,T]}$ is the time-reversed process of the standard Ornstein-Uhlenbeck process defined in (19) and corresponds to a Variance-Preserving diffusion model (Song et al., 2021). For any $t \in [0, T]$, given $X \sim \pi$, we therefore have

$$Y_t = \mathrm{e}^{-(T-t)}X + \sqrt{1 - \mathrm{e}^{-2(T-t)}}Z \ ,$$

where Z is distributed according to the standard centered Gaussian distribution. In this case, in a similar fashion to the SL posterior given in (8), the conditional density of X given $Y_t = y \in \mathbb{R}^d$ can be defined as

$$q_t(x|y) \propto \pi(x)\mathsf{N}(x; \mathrm{e}^{T-t}y, (\mathrm{e}^{2(T-t)} - 1)\,\mathrm{I}_d), \ t \in [0, T) \ ,$$

and the denoiser function as $u_t(y) = \int_{\mathbb{R}^d} x q_t(x|y)\mathrm{d}x$. Similarly to our framework, the random vector $u_t(Y_t)$ can be seen as the denoiser of Y_t. Following Tweedie's formula given in Lemma 4, the denoiser function is related to the score of p_{T-t} for any $t \in [0, T)$ by

$$\nabla \log p_{T-t}(y) = -\frac{y}{1 - \mathrm{e}^{-2(T-t)}} + \frac{\mathrm{e}^{-(T-t)}}{1 - \mathrm{e}^{-2(T-t)}}u_t(y) \ .$$

Therefore, the SDE (32) describing the denoising process is strictly equivalent to the following SDE

$$\mathrm{d}Y_t = \left\{\frac{\mathrm{e}^{-2(T-t)} + 1}{\mathrm{e}^{-2(T-t)} - 1}Y_t + \frac{2\mathrm{e}^{-(T-t)}}{1 - \mathrm{e}^{-2(T-t)}}u_t(Y_t)\right\}\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \ Y_0 \sim p_T \ . \tag{33}$$

This last SDE has to be compared with the SDE describing our observation process in (6): in both cases, the drift involves the denoiser function, which is not tractable in practice. Here, the target distribution is approximated by the distribution of Y_T, while it is approximated by the distribution of $Y_T^\alpha/\alpha(T)$ in our framework.

The approach of (Huang et al., 2024) to sample from the SDE (33), while handling the estimation of the denoiser u_t, is close to ours. Indeed, after discretizing this SDE with a certain time discretization $(t_k)_{k=1}^K$ of $[0, T]$, they propose to estimate the denoiser at time t_k with a Markov Chain Monte Carlo estimation. To do so, at each step t_k, they approximately sample from the random posterior $q_{t_k}(\cdot|\tilde{Y}_{t_k})$ with ULA (while we use MALA), where $\tilde{Y}_{t_k}$ is the k-th realization of their discretized process. Besides this, they also propose a Langevin-within-Langevin initialization to approximately sample from p_T, that involves the corresponding posterior q_0. Although they briefly discuss the difficulty of this initialization by considering the log-Sobolev constants of p_T and q_0, their analysis lacks readability and does not emphasize any crucial trade-off on T (which corresponds to our t_0).

Their scheme can nonetheless be analyzed as a localization scheme under the scope of what we call "duality of log-concavity", see Section 4.3. Assume that π verifies **A**0 and is not log-concave. Since we have $q_0(x|y) \propto \pi(x)\mathsf{N}(x; \mathrm{e}^T y, (\mathrm{e}^{2T} - 1)\,\mathrm{I}_d)$, one can show that

(a) if T is large: q_0 will be close to π (not log-concave), while p_T will be close to $\mathsf{N}(0, \mathrm{I}_d)$ (log-concave),

(b) if T is small: q_0 will localize as a Dirac mass (log-concave), while p_T will be close to π (not log-concave).

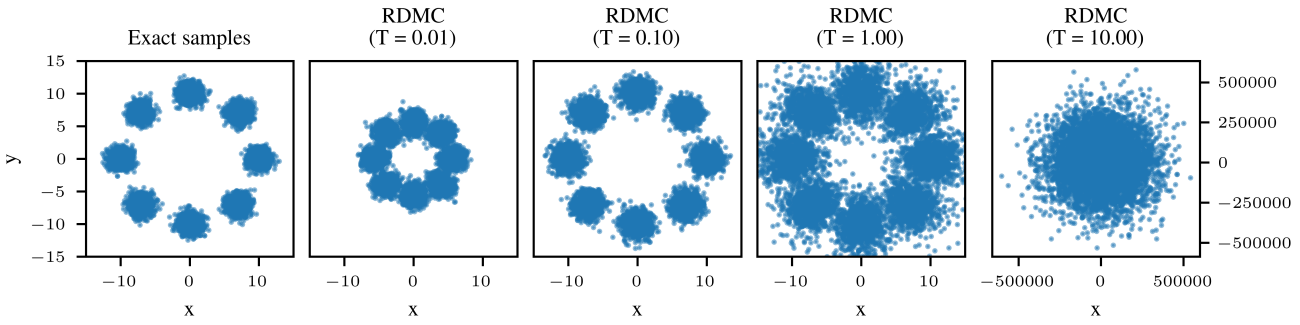

Figure 13: Sampling the 8-Gaussians distribution via RDMC with different values of T.

Therefore, the setting of T is crucial in RDMC to ensure a proper initialization via Langevin-within-Langevin algorithm. However, the authors claim that setting T large would only incur computational waste and that RDMC shows insensitivity to T, aside from the case where T is too close to 0[11]. As explained above, these claims are not realistic, since sampling from the posterior q_0 becomes as hard as sampling from π when T is large. We illustrate this fundamental limitation in Figure 13.

In contrast, we pay a particular attention to point out the importance of our hyper-parameter t_0 throughout our analysis, with practical and theoretical perspectives. Finally, we highlight that RDMC relies on a uniform time discretization (while we propose a SNR-adapted discretization), and do not use persistent initialization of their Markov chains in posterior sampling.

G.2. Multi-Noise Measurements sampling (Saremi & Srivastava, 2022; Saremi et al., 2023; 2024)

Given $M \geq 1$, the Multi-Noise Measurements (MNM) model introduced by (Saremi & Srivastava, 2022) defines a sequence of random variables $(Y^m)_{m=1}^M$ as

$$Y^m = X + \sigma Z^m \ ,$$

where $X \sim \pi$, $\sigma > 0$ and $(Z^m)_{m=1}^M$ are independently distributed according to the standard centered Gaussian distribution. This *non-Markovian* stochastic process can be seen as a denoising process. The rationale behind this formulation is that simulating an increasing number of *equally* noised measurements of a sample from π helps to obtain more information on this sample, and finally approximate it. We now explain how it can be interpreted as a non-Markovian analog to the standard localization scheme given in (1).

For any $m \in \{1, \dots, M\}$, the conditional density of X given the m-tuple $Y^{1:m} = y_{1:m} \in (\mathbb{R}^d)^m$ is defined as

$$q_m(x|y_{1:m}) \propto \pi(x)\mathsf{N}(x; \bar{y}_{1:m}, \sigma^2/m \, \mathrm{I}_d) \ ,$$

where $\bar{y}_{1:m}$ denotes the empirical mean of the noised measurements, *i.e.*, $\bar{y}_{1:m} = (1/m)\sum_{i=1}^m y_i$. By defining $u_m(y_{1:m}) = \int_{\mathbb{R}^d} x q_m(x|y_{1:m})\mathrm{d}x$, one can derive the Bayes estimator of X given $Y^{1:m}$ as the random vector $u_m(Y^{1:m})$ (Saremi & Srivastava, 2022). Interestingly, this denoiser shares the same rate of convergence as the denoiser of standard stochastic localization given in Proposition 8 when $t = m$. Indeed, (Saremi et al., 2024, Proposition 2) states that for any target distribution π, we have $W_2(\pi, \tilde{\pi}_m) \leq \sigma\sqrt{d/m}$, where $\tilde{\pi}_m$ denotes the distribution of $u_m(Y^{1:m})$. In other words, the MNM denoiser localizes to X with the same localization rate as the standard SL denoiser, defined in Section 2, by taking $T = M$.

To sample from π, the MNM approach consists in first sampling $Y^{1:M}$, and then computing the denoiser $u_M(Y^{1:M})$, in the same fashion as in the SL framework. Recently, (Saremi et al., 2024) proposed to tackle the sampling of the M-tuple $Y^{1:M}$ by first simulating Y^1 and then sequentially sampling Y^m given $Y^{1:m-1}$ for $m \in \{2, \dots, M\}$ using a Monte Carlo Markov Chain method - in this case, Underdamped Langevin Algorithm (Sachs et al., 2017). At step m, the Langevin procedure then involves pointwise evaluations of the conditional score of the distribution of Y^m given $Y^{1:m-1}$ (or simply the score of the distribution of Y_1 when $m = 1$). This introduces the *Once-At-a-Time* (OAT) algorithm. The interest of such method lies in the fact that, under Assumption **A**0 that may include multi-modal distributions, the distributions to sample from are increasingly log-concave with m, see (Saremi et al., 2024, Theorem 1). Hence, the most challenging step of this approach lies in the sampling of Y^1, in the same spirit as in `SLIPS`.

[11] See (Huang et al., 2024, Appendix F.4).

In their work, (Saremi et al., 2024) mainly consider the case where the scores are analytically available. To handle realistic settings, they implement an IS estimator, see (Saremi et al., 2024, Section 4.2.1), but their numerical results show that its performance significantly degrades compared to setting of perfect knowledge of the score, see (Saremi et al., 2024, Appendix H). We include this approach in our numerical benchmark. Alternatively, for high-dimensional settings, they propose an estimator based on posterior sampling, see (Saremi et al., 2024, Appendix F.2.), in the same fashion as in SLIPS. However, they do not pay attention to the limitation of this approach, in particular at the initialization of their algorithm. Indeed, while (Saremi et al., 2024, Theorem 1) suggests to take σ very large to ensure that the distribution of Y^1 is approximately log-concave (*i.e.*, taking t_0 small in SLIPS), the corresponding posterior becomes unfortunately closer to π, and then hard to sample with standard MCMC methods when π is not log-concave. This fundamental constraint is well illustrated in (Saremi et al., 2024, Figure 7, left), where the posterior sampling approach is shown to systematically fail for an arbitrary choice of σ, independently of the computational budget allocated to the MCMC sampling. Therefore, the combination of OAT with posterior sampling requires a trade-off on the hyper-parameter σ, similarly to t_0 in SLIPS, that reflects once again the "duality of log-concavity".

H. Details on numerical experiments

H.1. The failure of MCMC methods when targeting multi-modal distributions

As we show in Figure 14, *local* MCMC samplers such as MALA, HMC, NUTS or ESS tend to produce Markov chains that get trapped in modes while our methodology SLIPS generates samples reaching both modes.

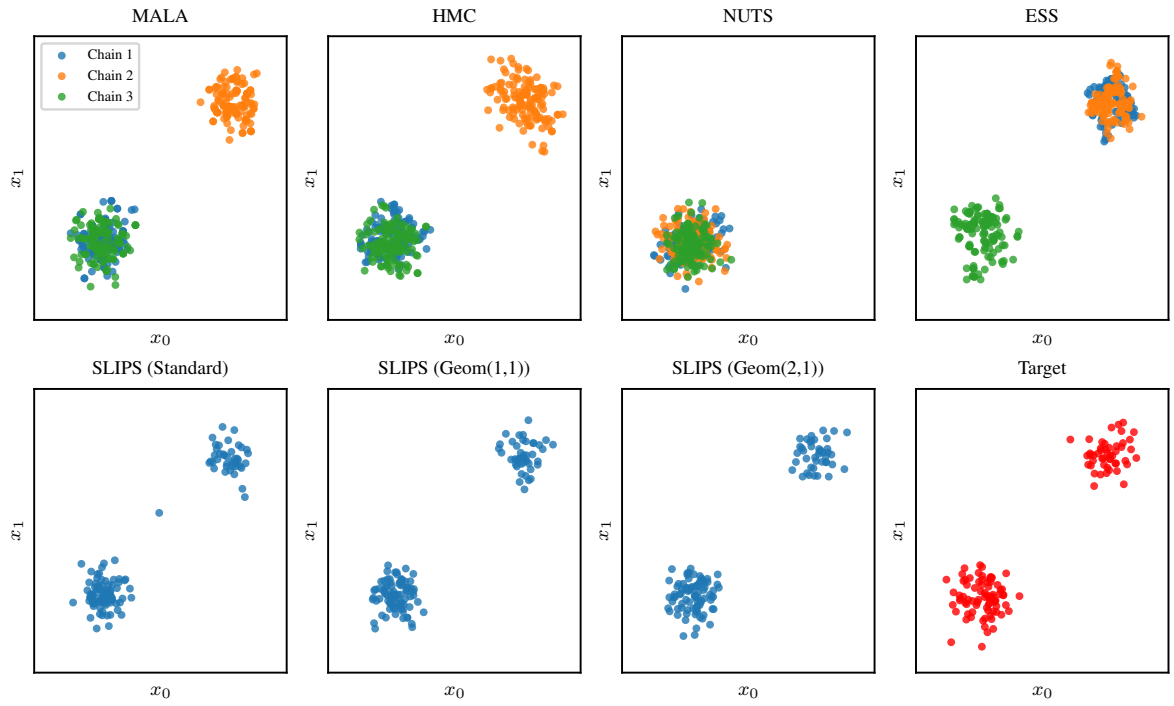

Figure 14: Samples from the Gaussian mixture defined in Section 6 ($d = 8$) obtained using different algorithms. Here, we display the first two coordinates of $n = 128$ samples. For SLIPS, we used the values of t_0 and η given in Table 5, $K = 10$ discretization steps and $n_{\text{MCMC}} = 16$ MCMC steps. On the other hand, the standard MCMC algorithms produce 3 chains, each one being of size $K \times n_{\text{MCMC}} \times n$ for fairness of comparison on the number of target density evaluations; we only display the n last MCMC samples of each chain. The chains were initialized in $\mathsf{N}(0, \sigma^2 \mathrm{I}_d)$. For HMC, we grid-searched an optimal trajectory length in $[4, 8, 12, 16]$ which ended up being 8 (the step-size was automatically tuned). For ESS, we used $\mathsf{N}(0, \sigma^2\, \mathrm{I}_d)$ as a prior.

Table 3: Hyper-parameter grids used for SLIPS on different targets.

		MIXTURE OF GAUSSIANS AND BAYESIAN	ϕ^4 MODEL	OTHERS
STANDARD	η	$\{5.0\}$	$\{5.7, 6.1\}$	$\{5.0, 5.7\}$
	t_0	$\{0.03, 0.05, 0.1, 0.2, 0.4\}$	$\{0.8, 1.0, 1.2, 1.4, 1.8\}$	$\{0.1, 0.2, 0.4, 1.0, 1.2\}$
GEOM(1,1)	η	$\{5.0\}$	$\{5.7, 6.1\}$	$\{4.6, 5.0\}$
	t_0	$\{0.03, 0.05, 0.1, 0.15, 0.25\}$	$\{0.30, 0.35, 0.40, 0.45\}$	$\{0.1, 0.15, 0.20\}$
GEOM(2,1)	η	$\{5.0\}$	$\{5.7, 6.1\}$	$\{4.6, 5.0\}$
	t_0	$\{0.15, 0.20, 0.25, 0.35, 0.45\}$	$\{0.40, 0.45, 0.50, 0.55\}$	$\{0.30, 0.35, 0.45\}$

H.2. Algorithms and hyper-parameters

Using the information from R_π and adapting the algorithms. As our definition of σ in SLIPS requires to have an estimate of R_π (see Section 3.2), we informed other algorithms with such knowledge when possible. More precisely, we use **A**0 and consider the approximation $R_\pi^2 = d(R^2 + \tau^2)$ where R and τ are assumed to be known.

- For SMC and AIS, we use a starting distribution ρ_0 as the centered Gaussian with variance $R^2 d\, \mathrm{I}_d$;
- For OAT, we use σ as per (Saremi et al., 2024, Equation 4.3) by setting $\sigma^2 = R^2 d - \tau^2$ (ensuring log-concavity);
- For RDMC, we do not include this information as the authors do not give any heuristic on T with respect to the variance of the target distribution.

Moreover, the implementations of the algorithms were slightly adjusted from their general definition. SMC and AIS use a MALA kernel (which leaves the target distribution invariant) as transition kernel. Additionally, RDMC also uses a MALA kernel for posterior sampling instead of ULA. This reduces the estimation bias and also enables automatic tuning of the Langevin step-size by leveraging the acceptance ratio (here we adapt the step-size to maintain the acceptance ratio at 75%). Still on RDMC, we drop the first 50% of the MCMC samples to ignore the warm-up period in the estimation. We also ignore the warm-up period with the same proportion in SLIPS. In SLIPS, we reuse the step-size from the MCMC sampling of q_{t_k} when sampling $q_{t_{k+1}}$ and we also initialize the chains of the later with the last samples of the chain from the former. This is an intuitive initialization when looking at the bottom row of Figure 3 as the modes of the posterior seem to be stable and q_{t_k} is expected to be close to $q_{t_{k+1}}$.

Estimating the scores in OAT. In OAT, denoting $p(y)$ the likelihood of the measurement process $Y = X + \sigma Z$ with $X \sim \pi$ and $Z \sim \mathsf{N}(0, \mathrm{I}_d)$, the score can be written using the following identity from (Saremi et al., 2024, Appendix F) : $\nabla \log p(y) = \sigma^{-1} \mathbb{E}_{Z \sim q(\cdot|y;\sigma)}[Z]$ where the posterior $q(\cdot|y)$ is defined by $q(z|y) \propto \pi(y + \sigma z)\mathsf{N}(z; 0, \mathrm{I}_d)$. This means that the score $\nabla \log p(y)$ can be estimated with the following IS estimator $\nabla \log p(y) \approx \sigma^{-1} \sum_{i=1}^N w_i Z_i$ where $Z_i \overset{i.i.d.}{\sim} \mathsf{N}(0, \mathrm{I}_d)$ and $w_i = \pi(y + \sigma Z_i) / \sum_{j=1}^N \pi(y + \sigma Z_j)$. We reduce the variance of this estimator by applying the antithetic trick.

Selecting the hyper-parameters of the algorithms. For each algorithm, we search its hyper-parameters within a predetermined grid. The selection is based on the metrics which will be later detailed. The metrics were computed by comparing 4096 samples against true samples. We globally fixed the computational budget by setting the SDE discretization of SLIPS as $K = 1024$. Moreover, we define the number of MCMC steps of SLIPS, denoted by L, to be equal to 32 except with the mixture of Gaussian in high dimensions where it is equal to 48, 64 and 96 in dimensions 32, 64 and 128 respectively and in the ϕ^4 experiments were it is set to 64. The selected hyper-parameters for each algorithm are summarized in Table 4. Below, we detail how the grids were built for each one.

- The SMC and AIS algorithms define a sequence of annealed distributions ρ_k for $k \in \{0, \dots, K\}$ from ρ_0 (defined above) to $\rho_K = \pi$ as $\rho_k \propto \exp((1 - \beta_k) \log \rho_0 + \beta_k \log \pi))$ where $(\beta_k)_{k=0}^K$ is a linear schedule of size K between 0 and 1. Both algorithms used $N = 4096$ particles;
- The RDMC algorithm has a single hyper-parameter T which we search within the grid $T \in \{-\log 0.99, -\log 0.95, -\log 0.9, -\log 0.8, -\log 0.7\}$ given in (Huang et al., 2024, Appendix F.1). However, we also find that large T may not work systematically (see Figure 13). We use 16 steps of (Huang et al., 2024, Algorithm 3) with 4 MCMC chains. The chains are initialized with an IS approximation of the posterior powered by 128 particles.

The SDE is discretized over K steps and the expectation leading to the drift is estimated with L Langevin steps. The initial sample is distributed according to $\mathsf{N}(0, (1 - \exp(-2T))\,\mathrm{I}_d)$, the best estimation of p_T that we have, and the initial step-size is taken according to its variance ;

- The OAT algorithm has its parameter M set as $\lfloor K/2 \rfloor$ and uses as many MCMC steps per noise level as `SLIPS` or RDMC. This choice of K ensures that the computational complexity of OAT is on par with the other algorithms. The Langevin steps are done using the underdamped Langevin algorithm as suggested by the authors. Its step-size is searched in $\{0.03, 1.0\}$ and the efficient friction is searched in $\{0.0625, 0.05\}$ as recommended by the authors. These prescriptions were extracted from (Saremi et al., 2024, Appendix G). The OAT algorithm has slightly shorter grid sizes because of its prohibitive computational cost;
- The `SLIPS` algorithm is declined in three flavors depending on the choice of the schedule α. Since the choice of t_0 is sensitive to the accuracy of the estimation of R_π (different values of R_π will shift the log-SNR upwards or downwards), we decided to search the hyper-parameters in different areas depending on the target distribution. Those grids can be found in Table 3. The values for t_0 were chosen by approximate equal log-SNR spacing in $[-3.5, -1.0]$ for Gaussian mixtures and Bayesian logistic regression, in $[-1.0, 0.2]$ for ϕ^4 and $[-2.0, 0.0]$ for the others. The values for η were chosen to be around 5.0.

Table 4: Hyper-parameters selected for the experiments for each algorithm and target density.

TARGET	RDMC	OAT		SLIPS STANDARD		SLIPS GEOM(1,1)		SLIPS GEOM(2,1)	
	TIME T	STEP-SIZE δ	FRICT. $\delta\gamma$	η	t_0	η	t_0	η	t_0
8-GAUSSIANS	$-\log(0.80)$	0.05	0.03	5.7	0.60	5.7	0.35	5.0	0.35
RINGS	$-\log(0.80)$	0.0625	1.0	4.6	1.20	4.6	0.10	4.6	0.30
FUNNEL	$-\log(0.90)$	0.05	0.03	5.0	1.00	4.6	0.30	4.6	0.40
MIXTURE ($d = 8$)	$-\log(0.70)$	0.0625	0.03	5.0	0.40	5.0	0.25	5.0	0.45
MIXTURE ($d = 16$)	$-\log(0.70)$	0.0625	0.03	5.0	0.20	5.0	0.15	5.0	0.35
MIXTURE ($d = 32$)	$-\log(0.70)$	0.05	0.03	5.0	0.10	5.0	0.10	5.0	0.25
MIXTURE ($d = 64$)	$-\log(0.70)$	0.0625	0.03	5.0	0.05	5.0	0.05	5.0	0.20
IONOSPHERE	$-\log(0.95)$	0.0625	0.03	5.0	0.03	5.0	0.03	5.0	0.15
SONAR	$-\log(0.95)$	0.0625	0.03	5.0	0.03	5.0	0.03	5.0	0.15
ϕ^4 ($b = 0$)	$-\log(0.95)$	0.0625	0.03	5.7	0.80	5.7	0.30	5.7	0.40
ϕ^4 ($b = 0.025$)	$-\log(0.95)$	0.0625	1.0	5.7	1.80	5.7	0.35	6.1	0.45
ϕ^4 ($b = 0.05$)	$-\log(0.70)$	0.05	1.0	6.1	1.00	5.7	0.30	5.7	0.40
ϕ^4 ($b = 0.075$)	$-\log(0.70)$	0.0625	0.03	5.7	1.80	5.7	0.35	5.7	0.40
ϕ^4 ($b = 0.1$)	$-\log(0.90)$	0.05	0.03	5.7	1.40	5.7	0.45	5.7	0.40

H.3. Target distributions and metrics

8 Gaussians, Rings and Funnel distributions and their respective metrics.

(a) The 8 Gaussians distribution consists of 8 equally weighted Gaussian distributions with mean $\mathbf{m}_i = 10 \times (\cos(2\pi i/8), \sin(2\pi i/8))$ for $i \in \{0, \dots, 7\}$ and covariance $0.7\,\mathrm{I}_2$. This distribution satisfies **A**0 with $R = 10/\sqrt{2}$ and $\tau^2 = 0.7$.

(b) The Rings distribution is the inverse polar reparameterization of a distribution p_z which has itself a decomposition into two univariate marginals p_r and p_θ: p_r is a mixture of 4 Gaussian distributions $\mathsf{N}(i+1, 0.15^2)$ with $i \in \{0, \dots, 3\}$ describing the radial position and p_θ is a uniform distribution over $[0, 2\pi]$, which describes the angular position of the samples. This distribution satisfies **A**0 with $R = 4/\sqrt{2}$ and $\tau = 0.15$.

(c) The Funnel distribution (Neal, 2003) has its density given by $x_{1:10} \mapsto \mathsf{N}(x_1; 0, \sigma^2 \mathrm{I}_{10})\mathsf{N}(x_{2:10}; 0, \exp(x_1)\mathrm{I}_{10})$ where $\sigma^2 = 9$. We approximate R by the maximum scalar standard deviation of the distribution, *i.e.*, $R = 2.12$, and set $\tau = 0$.

For 8 Gaussians and Rings, we evaluate the quality of sampling by approximating the entropy-regularized 2-Wasserstein distance defined in Appendix A, with regularization $\varepsilon = 0.05$ via POT library (Flamary et al., 2021). Regarding the Funnel, we evaluate the quality of sampling using the Kolmogorov-Smirnov distance, see Appendix A, and adopt the sliced version from (Grenioux et al., 2023b, Appendix D.1).

Bayesian Logistic Regression models. Consider a training dataset $\mathcal{D} = \{(x_j, y_j)\}_{j=1}^M$ where $x_j \in \mathbb{R}^d$ and $y_j \in \{0, 1\}$ for all $j \in \{1, \dots, M\}$. We evaluate the likelihood of a pair (x, y) as given by $p(y|x; w, b) = \mathrm{Bernoulli}(y; \sigma(x^T x + b))$ where $w \in \mathbb{R}^d$ is a weight vector, $b \in \mathbb{R}$ is an intercept and σ is the sigmoid function. Given a prior distribution $p(w, b)$, we sample from the posterior distribution $p(w, b|\mathcal{D}) \propto p(\mathcal{D}|w, b)p(w, b) = \prod_{j=1}^M p(y_j|x_j; w, b)p(w, b)$. The prior is built as $p(w, b) = \mathsf{N}(w; 0, \mathrm{I}_d)\mathsf{N}(b; 0, (2.5)^2)$. In our experiments, we approximate R by the maximum scalar standard deviation of the prior distribution, *i.e.*, $R = 2.5/\sqrt{d+1}$ and set $\tau = 0$. The quality of the samples are obtained by computing the mean predictive log-likelihood (i.e., computing $p(w, b|\mathcal{D}_{\text{test}})$ with $\mathcal{D}_{\text{test}}$ a separate test dataset).

High dimension Gaussian mixture. We consider the mixture of two Gaussian distributions defined in (24). Following the computations of Appendix C, we set $R = 4/3$ and $\tau^2 = 0.05$. Complementary to the estimation error on the relative weight given in the main paper, we also report the Sliced Wasserstein distance (Bonneel et al., 2015; Nadjahi et al., 2019) on Figure 15 (*left*). This figure shows that `SLIPS` also recovers the local structure of the target distribution quite well. Moreover, Figure 15 (*right*) shows that SMC and AIS are unable to scale with dimension and collapse into a single mode.

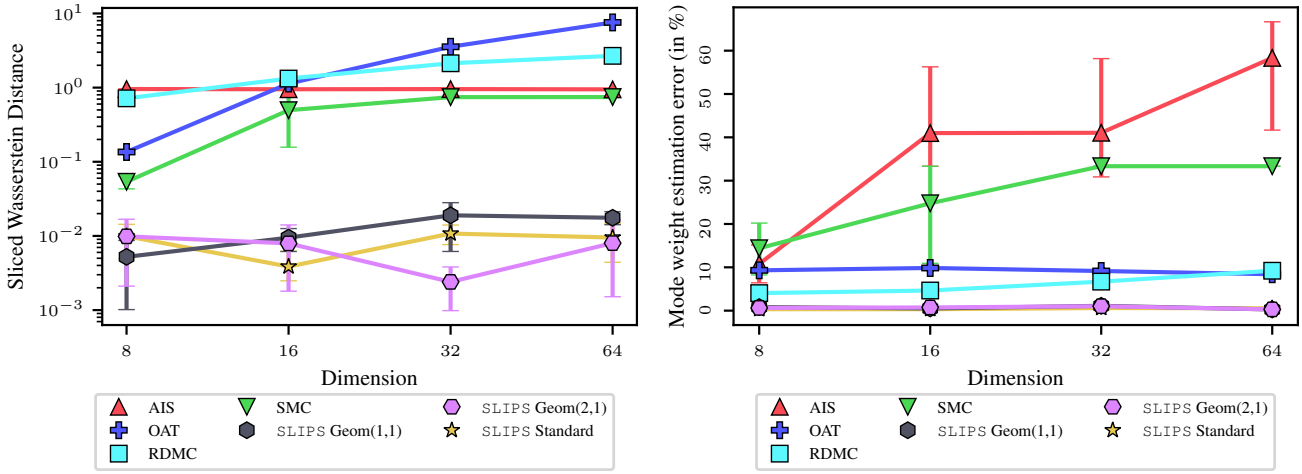

Figure 15: **Left**: Sliced Wasserstein distance computed on bimodal Gaussian mixtures with increasing dimension. **Right**: Estimation error when computing the relative weight of the corresponding two modes.

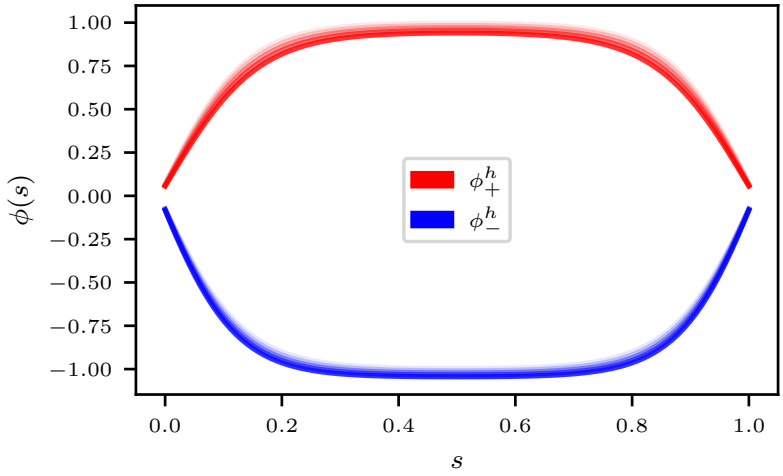

Figure 16: Modes of the ϕ^4 distribution (ϕ_+^h in shades of red and ϕ_-^h in shades of blue) for different values of h (corresponding to the transparency levels). The time interval $[0, 1]$ on the x-axis is discretized into a grid of size $d = 100$, which corresponds to the dimensionality of the samples.

The ϕ^4 distribution. The ϕ^4 model is a toy model defined as a continuous relaxation of the Ising model that serves the study of phase transitions in statistical mechanics. Following (Gabrié et al., 2022), we consider a version of the model discretized on a 1-dimensional grid of size $d = 100$. One configuration is therefore a d-dimensional vector $(\phi_i)_{i=1}^d$. We additionally clip the field to 0 at both extremities by defining the extra $\phi_0 = \phi_{d+1} = 0$. The negative log-density of the distribution writes

$$\ln \pi_h(\phi) = -\beta \left(\frac{ad}{2} \sum_{i=1}^{d+1} (\phi_i - \phi_{i-1})^2 + \frac{1}{4ad} \sum_{i=1}^{d} (1 - \phi_i^2)^2 + h\phi_i \right) . \tag{34}$$

We chose parameter values for which the system is bimodal, $a = 0.1$ and inverse temperature $\beta = 20$, and vary the value of h. We denote by w_+ the statistical occurrence of configurations such that $\phi_{d/2} > 0$ and w_- the statistical occurrence of configurations such that $\phi_{d/2} < 0$. At $h = 0$, the measure is invariant under the symmetry $\phi \to -\phi$, such that we expect $w_+ = w_-$. For $h > 0$, the negative mode dominates. We plot the two modes on Figure 16.

When d is large, the relative probability of the modes can be estimated by a Laplace approximations at 0-th and 2-nd order. Denoting by ϕ_+^h and ϕ_-^h the local maxima of (34), these approximations yield respectively,

$$\frac{w_-}{w_+} \approx \frac{\pi_h(\phi_-^h)}{\pi_h(\phi_+^h)} , \quad \frac{w_-}{w_+} \approx \frac{\pi_h(\phi_-^h) \times |\det H_h(\phi_-^h)|^{-1/2}}{\pi_h(\phi_+^h) \times |\det H_h(\phi_+^h)|^{-1/2}} ,$$

where H_h is the Hessian of the function $\phi \to \ln \pi_h(\phi)$. In our experiments, we considered $R = 4.5$ and $\tau = 10^{-2}$ by running MALA chains started in each mode : τ^2 is set as the variance within the modes while R corresponds to the distance between the modes and 0_{100}. In this setting, we observed that SMC and AIS suffered from mode collapse, while OAT and RDMC produced degenerate samples; in contrast, our methodology `SLIPS` recovers accurate samples from the target distribution, with correct relative weight, as shown in Figure 5.

H.4. Empirical complexity of `SLIPS`

The results of Figure 17 ($d = 32$) show that under a much lower computational budget ($K = 20$ here) than in Section 6 ($K = 1024$), `SLIPS` maintains the same performance, while still outperforming its competitors. Moreover, these results demonstrate that `SLIPS` has very reasonable execution times (below 1 minute to obtain 8192 samples) making it completely practical. Lastly, Figure 17 emphasizes that, even using high computational budgets, competitors cannot solve such multi-modal tasks in high-dimension. Note that those observations stay valid in problems with lower dimension (see Figure 18 where $d = 4$).

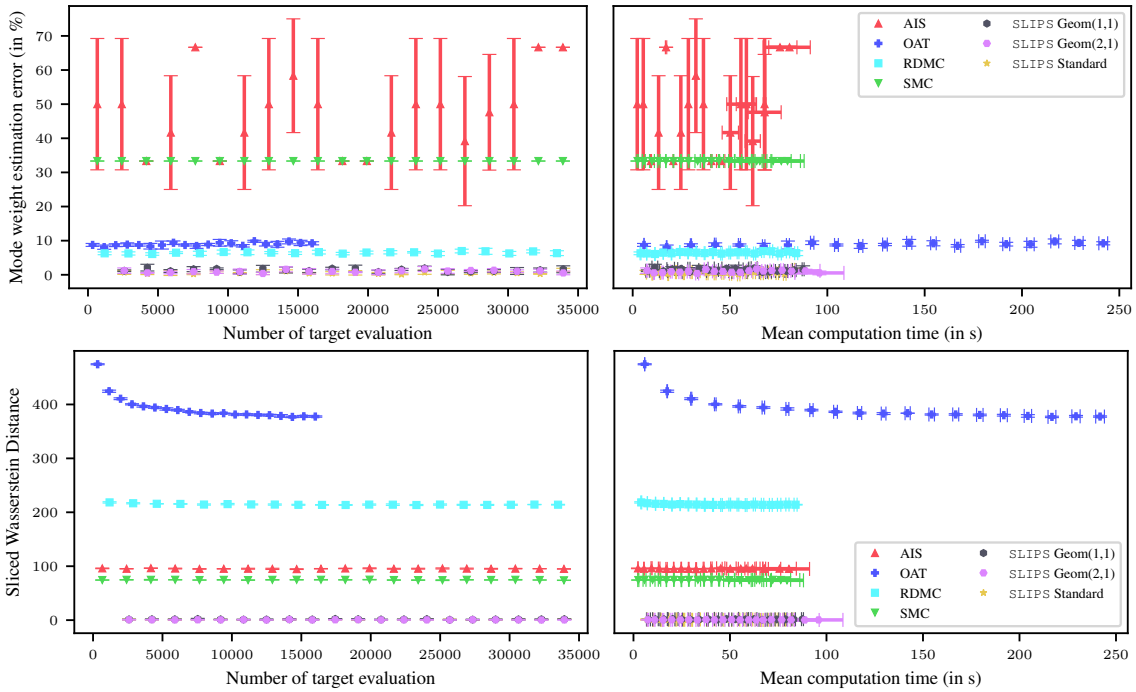

Figure 17: Metrics when sampling from the Gaussian mixture defined in Section 6 ($d = 32$) using different algorithms. **Top**: Weight estimation error. **Bottom**: Sliced Wasserstein Distance. **Left**: Metric depending on the number of target density evaluations. **Right**: Metric depending on wall time. The computational budgets are computed by evolving K linearly from 20 to 90. The number of MCMC steps is fixed at 32. The computations were run on the same Nvidia V100 GPU.

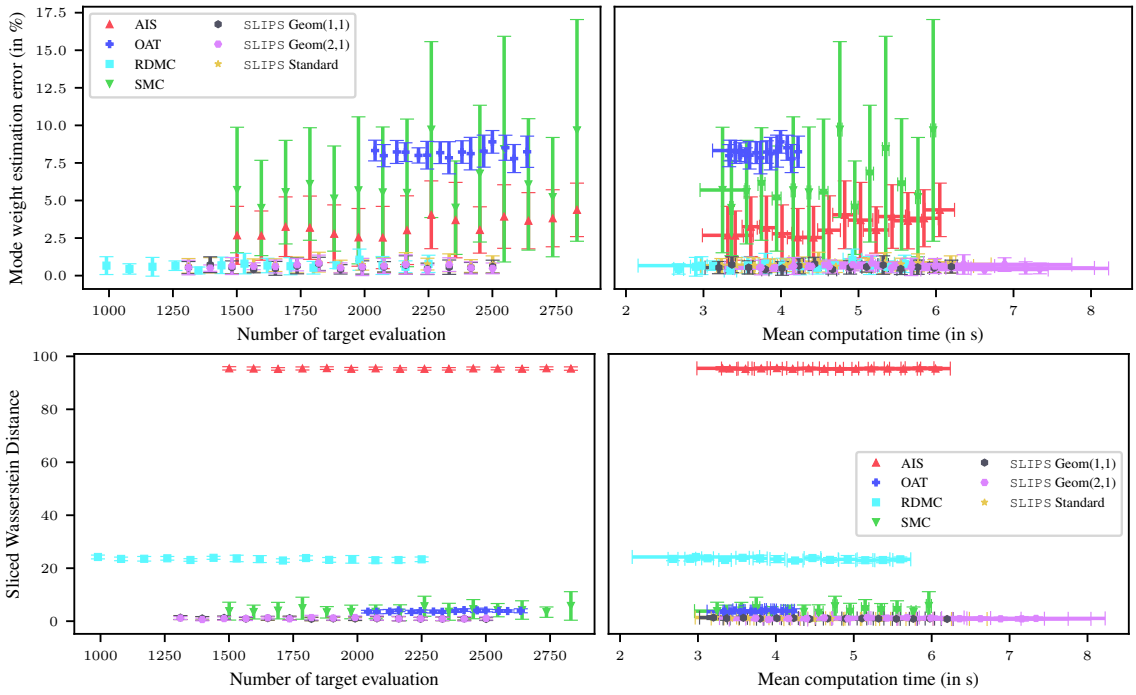

Figure 18: Metrics when sampling from the Gaussian mixture defined in Section 6 ($d = 4$) using different algorithms. **Top**: Weight estimation error. **Bottom**: Sliced Wasserstein Distance. **Left**: Metric depending on the number of target density evaluations. **Right**: Metric depending on wall time. The computational budgets are computed by evolving K linearly from 20 to 90. The number of MCMC steps is fixed at 16. The computations were run on the same Nvidia V100 GPU. The computational budgets were aligned to match SLIPS's.