IMPROVING THE EVALUATION OF SAMPLERS ON MULTI-MODAL TARGETS

Louis Grenioux* **& Maxence Noble***
CMAP, CNRS
École polytechnique, Institut Polytechnique de Paris
91120 Palaiseau, France
{louis.grenioux,maxence.noble-bourillot}@polytechnique.edu

Marylou Gabrié
Laboratoire de Physique de l'École normale supérieure
ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris
75005 Paris, France
marylou.gabrie@ens.fr

ABSTRACT

Addressing multi-modality constitutes one of the major challenges of sampling. In this reflection paper, we advocate for a more systematic evaluation of samplers towards two sources of difficulty that are mode separation and dimension. For this, we propose a synthetic experimental setting that we illustrate on a selection of samplers, focusing on the challenging criterion of recovery of the mode relative importance. These evaluations are crucial to diagnose the potential of samplers to handle multi-modality and therefore to drive progress in the field.

1 INTRODUCTION

Consider the problem of sampling from a probability density known up to a normalizing constant. More precisely, let π be a target distribution with probability density given by $x \mapsto \gamma(x)/\mathcal{Z}$, where $\gamma : \mathbb{R}^d \to \mathbb{R}_+$ can be pointwise evaluated and the normalizing constant $\mathcal{Z} = \int_{\mathbb{R}^d} \gamma(x)\mathrm{d}x$ is intractable. This reflection paper is concerned with cases where the distribution π is multi-modal, i.e., γ admits several local maxima. Such high-dimensional multi-modal distributions appear in various fields. In physics, they describe the thermodynamics of metastable systems, or the action of quantum field theories (Wolff, 1990). Bayesian modeling can also face multi-modal posterior distributions in cases where the observations fail to lift ambiguities between several distinctive sets of parameters (Yao et al., 2022). Hence, continuing efforts have been made to design sampling strategies that handle multi-modality. However, sampling from multi-modal distributions is notoriously difficult, especially in large dimension d.

There are two major sources of difficulties to overcome: (i) locating the modes, that is the different high-probability regions of π induced by its local maxima, and (ii) estimating their relative importance. Performing an exhaustive search of modes is equivalent to the problem of finding the local maxima of the possibly high-dimensional function γ, which is NP-hard Pardalos & Schnitger (1988). Although a flurry of heuristics have been developed to explore probability landscapes and find modes (Fletcher, 2000; Nocedal & Wright, 2006), this fact sets the problem of sampling from multi-modal distributions as arbitrarily hard in general.

In fact, even when the location of the modes is known, the second difficulty of estimating their relative proportion remains a significant challenge to this day. The ability of samplers to represent the different modes with correct proportions is, however, seldom reported, with exceptions including Grenioux et al. (2023); Blessing et al. (2024); Shi et al. (2024); Noble et al. (2025).

*Both authors contributed equally.

In this reflection paper, we wish to advocate for:

• Running systematic experiments on toy mixture distributions with increasing dimension and mode separation, as these two parameters are commonly mediating the hardness of the task for common sampling technics.

• Evaluation based on the criterion of accurate estimation of the statistical weight of each mode, possibly assuming knowledge of mode locations. The reasons are twofold: first because it is an interpretable metric, incurring almost no additional computational cost, and second precisely because it is a challenging test.

Running these experiments and evaluation would enable diagnosing strengths and weaknesses of newly proposed algorithms, which is crucial to foster the development of robust algorithms in the multi-modal setting, and therefore drive progress in the field.

In the following, we start with some reminders in Section 2 on the difficulty of sampling from multi-modal distributions, and in particular in estimating relative importance of modes. In Section 3, we examine the strengths and weaknesses of common sampling strategies in multi-modal scenarios and illustrate the importance of reporting performance within the proposed experimental framework.

2 What is hard in sampling from multi-modal distributions?

As mentioned in the introduction, the first challenge in sampling from a multi-modal distribution is identifying all its modes. We do not discuss this issue further, as exhaustive searches can be arbitrarily complex. Moreover, even when mode locations are known, accurate sampling remains difficult and practically relevant. In some cases, heuristics or expert knowledge can help identify modes. For instance, in physics and drug design (Jorgensen, 2004), determining molecular binding events involves two modes: the bonded complex and the unbound molecules.

Benefits and pitfalls of local samplers. If the target's mode locations are known, it may seem that the sampling problem is solved by using this prior information to initialize a Markov Chain Monte Carlo (MCMC) sampler. MCMCs based on local updates are among the most widely used sampling approaches, including Random Walk Metropolis-Hastings (RWMH) (Metropolis et al., 1953), Metropolis-Adjusted Langevin Algorithm (MALA) (Roberts & Tweedie, 1996), Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Brooks et al., 2011), and No U-Turn Sampler (NUTS) (Hoffman & Gelman, 2014). These algorithms have demonstrated both theoretical and practical efficiency when the target density satisfies log-concavity assumptions (Dalalyan, 2017; Durmus & Moulines, 2019; Dwivedi et al., 2018; Bou-Rabee et al., 2020). Paradoxically, the very properties that make these samplers effective for log-concave distributions hinder their performance in the multi-modal case, where log-concavity does not hold. While MCMC transition kernels are designed to guide the chain toward target modes, they also, with high probability, trap it within a mode, preventing broader exploration. Consequently, initializing an MCMC sampler at the modes of π—where γ is locally log-concave—may yield locally accurate samples while failing to capture the global mode weights of π.

Mode weight estimation. To define the mode weights of the distribution of interest π, one can leverage a partition of its support (taken equal to $\mathbb{R}^d$ here for simplicity) that draws a one-to-one correspondence between the target modes and space subsets. Then, the mode weights of π can be defined as the relative statistical weights of the constituents of this partition with respect to π. Formally, assume that π has K modes and that we have access to a partition $\{S_1, \dots, S_K\}$ of $\mathbb{R}^d$, where $S_k \subset \mathbb{R}^d$ represents a region including the k-th mode of π. The (un-normalized) k-th mode weight of π is simply defined by

$$w_k = \int_{\mathbb{R}^d} \mathbb{1}_{S_k}(x)\gamma(x)\mathrm{d}x = \mathcal{Z}\pi(S_k) \ . \tag{1}$$

The important quantities are the weight ratios $w_k/w_{k'}$ for $k' \neq k$ for which the normalizing constant $\mathcal{Z}$ simplifies. Therefore, the weights w_k may be defined up to a common multiplicative constant. Fixing the multiplicative constant, w_k can be interpreted as the normalization constant of the target distribution π restricted to S_k, with un-normalized density given by $x \mapsto \mathbb{1}_{S_k}(x)\gamma(x)$. As such, it becomes clear that mode weight estimation is related to the estimation of a normalization constant.

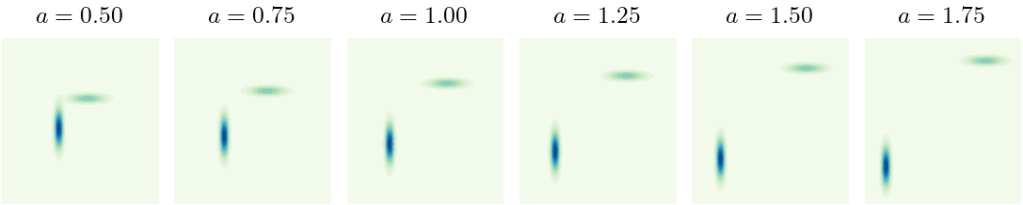

Figure 1: Density of the bi-modal Gaussian mixture used in our experiments ($d = 2$) with increasing values of a. As expected, a larger value of a makes the modes further from each other, which reinforces the sampling difficulty.

The need to go beyond local MCMC samplers for multi-modal distributions is well established, with annealing-based methods playing a key role (detailed in the next section). However, accurately estimating mode weights, even when mode locations are known, remains an open challenge in high dimensions. This issue parallels free energy computations in statistical mechanics, an active area of research (Lelièvre et al., 2010; Chipot, 2023).

Despite this, new samplers rarely evaluate their accuracy in estimating mode weights. Yet, this criterion complements common choices like integral probability metrics (IPMs) – such as the Wasserstein distances (Peyré & Cuturi, 2019) or the maximum mean discrepancy (Gretton et al., 2012) – the estimation of which is complex and typically comes without guarantees like unbiasedness. Mode weights, in contrast, can be easily estimated from the proportion of samples in each mode, offering a more interpretable measure that directly highlights the core challenge of multi-modal sampling.

3 Towards a systematic experiment for evaluating samplers

We propose a simple but challenging experimental setting to assess the ability of a given sampler at handling multi-modality by probing systematically two sources of difficulty: dimension and separation of the modes. We illustrate this setting by applying it to a set of common samplers, revealing their strengths and weaknesses in the task of mode weight estimation.

3.1 Description of the experimental setting

Multi-modal target distribution. We consider π defined as a d-dimensional bi-modal Gaussian mixture with unbalanced weights and covariances with strong components in different directions. More precisely, the mixture components are $q_1(x) = \mathrm{N}(x; -a\mathbf{1}_d, \Sigma_1)$, with weight $w = 2/3$, and $q_2(x) = \mathrm{N}(x; a\mathbf{1}_d, \Sigma_2)$, with weight $1 - w = 1/3$, where $a \in (0, \infty)$ controls the distance between the local maxima of π as showed in Figure 1, $\mathbf{1}_d$ is the d-dimensional vector with all components equal to 1, Σ_1, Σ_2 are fixed diagonal positive matrices with conditioning number equal to 20. To make this target even more challenging, we set those covariance matrices such that q_1 has marginally large variance in directions where q_2 has small variance (and inversely), see Appendix B for more details. In this setting, the mode partition of π is taken as $S_1 = \{x \in \mathbb{R}^d : q_1(x) > q_2(x)\}$ and $S_2 = S_1^c$, such that the induced w_1, defined in (1), verifies $w_1 \approx w$.

Mode weight metric. Below, we report the bias and variance for the Monte Carlo estimations of w_1, while varying two parameters: the distance between the mode centroids, controlled by the parameter a, and the dimension d. The results are displayed on Figures 2 and 3 for a selection of sampling methods. We provide, in Appendix B, a brief description each method with its hyperparameters and, below, comment on its performance. Small bias and variance indicate that the considered sampler is able to recover the mode proportions of our challenging test target both accurately and robustly, and is therefore a promising approach.

General remarks on the methodology. We emphasize that our numerical results should not be interpreted as a precise or definitive quantitative comparison between samplers for several reasons. First, their computational budgets were only roughly matched, see Figure 5 in Appendix B.

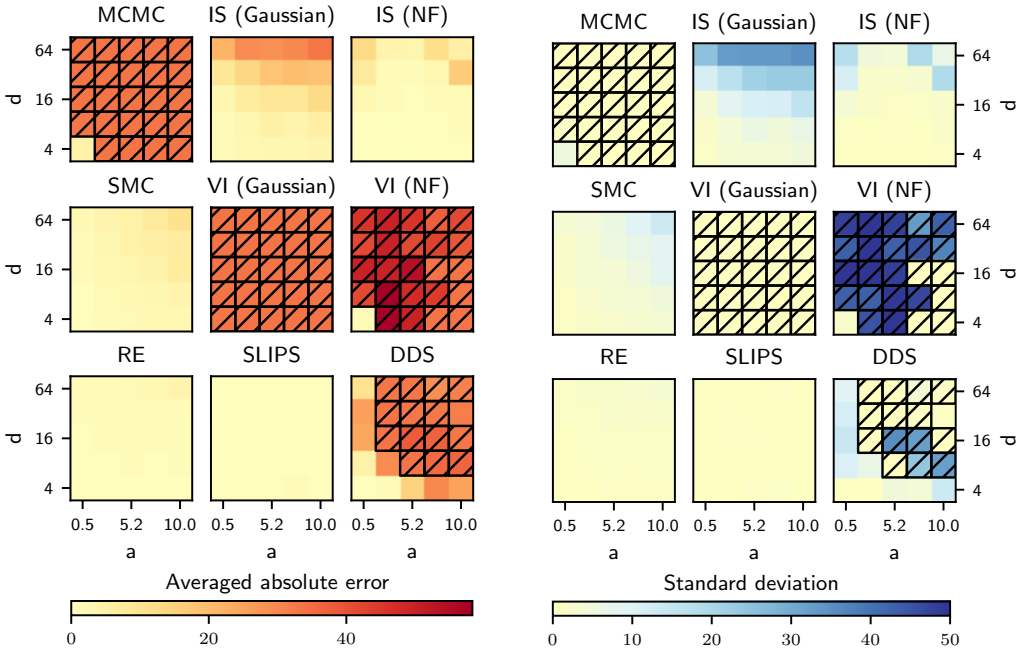

Figure 2: Results on mode weight estimation for the bi-modal Gaussian mixture defined in Section 3.1, when varying hyperparameters a and d. For each setting, we aggregate 48 Monte Carlo estimations, each one being computed over 8192 samples. Hashed areas indicate settings with systematic mode collapse in the sampling process. **(Left):** Averaged absolute error of the estimation with respect to the true mode weight $w \approx 66.7\%$. **(Right):** Standard deviation of the estimation.

Secondly, to showcase their best-case performance, their hyperparameters were individually tuned based on access to ground-truth samples, which would not be available in real-world scenarios, thus making their setting unrealistic for practical deployment.

Additionally, while the experimental target is admittedly a simplified, toy example compared to the complex multi-modal distributions discussed above, we note that most of the considered samplers already fail to perform well on this basic benchmark. This is a key observation: we argue that if a method cannot reliably sample from this controlled bi-modal setting, it is unlikely to succeed in more challenging, high-dimensional applications. Rather than offering fine-grained comparisons, these results are therefore intended to provide insight into the typical behavior and limitations of popular sampling algorithms.

3.2 Evaluation on various samplers

Local Markov Chain Monte Carlo. MCMC methods produce Markov chains that admit π as their stationary measure. Standard designs of transition kernels for log-concave distributions are based on discretizations of either Langevin or Hamiltonian dynamics. However, as already discussed in Section 2, the resulting sampling algorithms are inefficient when facing multi-modal targets, as moving between distinct modes becomes an unlikely event.

Testing an implementation of MALA, this behavior is easily evidenced in our experiment. The chains get trapped in one of the modes as soon as there exists a significant log-probability "barrier" between the modes, i.e., when a or d increases. The zero variance observed in these cases is not to be mistaken as a positive result. As the chains stay trapped in the mode they are initialized in, a setting we refer to as *mode collapsed*, the estimation of w_1 is consistently biased in the same manner.

Importance Sampling. Importance Sampling (IS) builds Monte Carlo estimators as re-weighted averages of samples from a tractable proposal distribution. The variance of IS estimations, related to the degeneracy of the so-called "importance weights", depends on the agreement between the target and the proposal distributions and is known to increase exponentially with d (Agapiou et al., 2017).

In Figure 2, the bias and variance of the estimator of w_1 are reported for two versions of IS. First, using a unimodal Gaussian proposal, the accuracy decreases rapidly along both the mode separation and dimension axes. As either of these parameters increases, the intersection of the typical sets of the target and proposal distribution decreases, leading to more degenerate weights. Using a normalizing flow has been proposed to adapt the proposal to the target and mitigate the issue (Müller et al., 2019; Noé et al., 2019). Here, the normalizing flow was trained on target samples to serve as a proposal, and indeed mitigates the degradation of performance, in particular with respect to the mode separation. The normalizing flow can learn to focus mass on different modes even if distant. However, the lack of perfect agreements between the target and the proposal starts showing as the dimension increases.

Variational Inference. Variational Inference (VI) reformulates the sampling problem as an optimization problem over the space of probability distributions. It seeks the closest distribution to the target within a family of tractable parametric distributions, known as the variational family. This problem requires a notion of discrepancy between probability distributions, which is chosen as the reverse Kullback-Leibler (KL) divergence in most cases. For multi-modal target distributions, many works have showed that even for variational family able to represent multiple modes, the VI procedure often leads to mode collapse, due to the mode-seeking behavior of the reverse KL divergence (MacKay, 2001; Jerfel et al., 2021; Blessing et al., 2024; Soletskyi et al., 2024).

In our experiment, we consider two families of variational distributions, unimodal Gaussian distributions and normalizing flows Rezende & Mohamed (2015). In the Gaussian case, we systematically observe a mode collapsed estimation, with the same consistent bias as in the MCMC setting, resulting in a zero-variance outcome. While using normalizing flows remediates the issue when the modes are almost connected and the dimension is low (lowest values of a and d), mode collapse is systematic beyond this setting.

Annealed sampling methods. Annealed samplers were developed to address the challenges posed by multi-modal distributions. The core idea is to introduce a sequence of distributions that gradually transition from an easy-to-sample distribution to the target distribution. These samplers leverage this sequence by breaking the original sampling problem into multiple, more manageable subproblems. In practice, this is often achieved through a geometric interpolation of densities.

SEQUENTIAL ANNEALED SAMPLING. A naive annealing approach consists in performing recursive MCMC sampling of the bridging distributions. Beginning with the simplest distribution and sequentially moving towards the target, each intermediate distributions is targeted by a standard MCMC sampler, that uses samples obtained for the preceding distribution as a warm-start. The Sequential Monte Carlo (SMC) algorithm (Del Moral et al., 2006) improves upon this approach by reweighting warm-start samples at certain steps using importance weights updated throughout the process. While SMC significantly outperforms the earlier samplers in handling multi-modality, it has been shown to degrade in high dimensions. Similar to Importance Sampling (IS), its resampling weights suffer from degeneracy (Li et al., 2014), a problem that can be exacerbated by the default geometric interpolation path (Máté & Fleuret, 2023). Increasing the number of intermediate steps can help mitigate this issue in higher dimensions but comes at a significant computational cost.

For our experiment, we leverage an optimal sequence of intermediate distributions based on the recommendations of Chopin et al. (2020), placing SMC in its most favorable setting. In Figure 2, we observe that SMC is perfectly able to recover the true mode proportion with high accuracy and robustness, confirming the superiority of annealing schemes over standard MCMC methods for multi-modal distributions. However, as expected, SMC performance deteriorates as soon as both of the mode separation and the dimension increase together as shown in Figure 3.

PARALLEL ANNEALED SAMPLING. A second approach amounts to perform MCMC sampling of each intermediate distribution in parallel, and to exchange states from the generated adjacent chains based on a Metropolis-Hastings acceptance step. This algorithm is often referred to as Replica Exchange (RE) (Swendsen & Wang, 1986) or Parallel Tempering (PT) (Geyer, 1991; Hukushima & Nemoto, 1996). Similarly to sequential annealing, this approach is known to suffer from high dimension, due to the degeneracy of the acceptance probabilities.

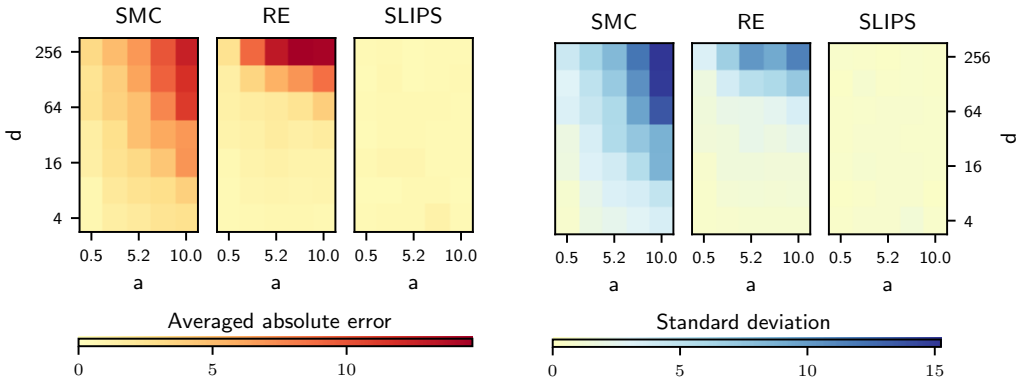

Figure 3: Same as Figure 2 for the best performing SMC, RE and SLIPS reaching higher-dimensions. For sake of pedagogy, the color scale is different from above (a consistent color scale can be found in Figure 4 of Appendix B). **(Left):** Averaged absolute error of the estimation with respect to the true mode weight $w \approx 66.7\%$. **(Right):** Standard deviation of the estimation.

Here, we test an implementation of RE with a number of intermediate distributions that is fixed over the choices of a and d [1]. Although RE accurately recovers the true mode proportion in the settings from Figure 2, its performance declines with increasing mode separation and dimension, as showed in Figure 3. This degradation appears slower than SMC, but note the adaptive SMC converged in a computational budget 20 times faster than the one considered for RE.

Diffusion-based sampling methods. Building on classical sequential annealing, the recently introduced diffusion-based sampling framework generates samples from the target distribution by following a Stochastic Differential Equation (SDE) that transitions from an easy-to-sample distribution to the target via a predefined stochastic interpolant process. This approach is inspired by the success of score-based diffusion models in generative modeling (Song et al., 2021; Albergo et al., 2023), where the goal is to generate samples without access to an explicit density. In practice, the drift term of this generative SDE is intractable and must be estimated, typically using Monte Carlo methods or variational approximations.

MONTE-CARLO DIFFUSION-BASED SAMPLING. Leveraging the fact that the intractable drift can be expressed as a conditional expectation, several recent works have proposed estimating it on-the-fly—i.e., while running the generative SDE—using unparameterized Monte Carlo methods (Huang et al., 2024a;b; He et al., 2024). A related approach, Stochastic Localization via Iterative Posterior Sampling (SLIPS) (Grenioux et al., 2024), estimates the drift using standard MCMC techniques. While these methods have demonstrated strong empirical performance, they require careful hyper-parameter tuning which becomes particularly challenging when no ground truth samples are available for their calibration, which is often the case in practice. This represents a critical limitation for their real-world applicability.

With this informed tuning, SLIPS performs very well at the mode weight estimation task as presented in Figure 2 and even outperforms SMC and RE in higher dimension, as displayed in Figure 3. This result may be explained by the fact that the sequence of intermediate densities induced by the diffusion path (which are also bi-modal) preserve the mode weights of the target distribution. This is contrary to the geometric interpolation paths (Máté & Fleuret, 2023) and avoids the use of advanced resampling schemes that notoriously fail when a or d increases.

VARIATIONAL DIFFUSION-BASED SAMPLING. Another line of research has proposed to learn the drift of the generative SDE with a neural network, by solving a Variational Inference problem, this time defined on the space of path measures (Zhang & Chen, 2022; Berner et al., 2023; Vargas et al., 2024; Chen et al., 2025). In this case, one aims to minimize a discrepancy between the generative SDE and a variational family of neural SDEs, i.e., SDEs whose drift is a neural network.

[1] We explored various types of discretization schedules for the intermediate distributions and observed that a logarithmic spacing was the best suited for this target.

In our experiment, we considered the Denoising Diffusion Sampler (DDS) (Vargas et al., 2023). The interest of such parameterized methods lies in their ability to amortize inference; once the drift is learned, samples can be obtained by solving the corresponding neural SDE. Their performance were found to severely depends on the values of their hyperaparameters Noble et al. (2025).

In Figure 2, we display results of DDS, where its hyperparameters were tuned favorably with respect to π, see Appendix B. Despite relative good performance in the most simple multi-modal settings, we still observe that it suffers from the joint increase of a and d. The difference with the Monte-Carlo diffusion-based sampler (SLIPS) may be explained by a manifestation of mode collapse in this extended VI problem.

4 PERSPECTIVES

We have demonstrated that evaluating a sampler's ability to recover mode weights accurately provides an interpretable benchmark for designing samplers for multi-modal distributions. Gradually increasing both dimensionality and mode separation in a controlled experiment reveals the breaking points of different samplers and helps identify the most promising approaches. Designed to handle multi-modal distributions, traditional annealing-based samplers, such as SMC and RE, are generally more robust than direct sampling methods. Recently introduced diffusion-based samplers also show promise, but their precise tuning and scalability remain open challenges.

While we focused on dimension and mode separation, the number of modes also adds to the challenge of sampling. The proposed experimental setting can be straightforwardly extended in this third axis of difficulty. In this case, the mode weight criterion translates into a comparison of the empirical histogram of visited modes with the ground-truth components' weights (see e.g., Grenioux et al. (2023)). We also note that the proposed framework does not include heavy-tails and complex local geometries, and it would be an interesting direction to define a systematic benchmark including also these features.

ACKNOWLEDGMENTS

We would like to thank Alain Durmus for many stimulating discussions and for reviewing the manuscript prior to submission. LG acknowledge funding from Hi! Paris. This work was partially performed using HPC resources from GENCI–IDRIS (AD011015234).

REFERENCES

S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017. ISSN 08834237, 21688745. URL http://www.jstor.org/stable/26408299.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

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

Denis Blessing, Xiaogang Jia, Johannes Esslinger, Francisco Vargas, and Gerhard Neumann. Beyond ELBOs: A large-scale evaluation of variational methods for sampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4205–4229. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/blessing24a.html.

Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *The Annals of Applied Probability*, 30(3):1209 – 1250, 2020. doi: 10.1214/19-AAP1528. URL https://doi.org/10.1214/19-AAP1528.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, May 2011. ISBN 9780429138508. doi: 10.1201/b10905. URL http://dx.doi.org/10.1201/b10905.

Junhua Chen, Lorenz Richter, Julius Berner, Denis Blessing, Gerhard Neumann, and Anima Anandkumar. Sequential controlled Langevin diffusions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=dImD2sgy86.

Christophe Chipot. Free Energy Methods for the Description of Molecular Processes. *Annual Review of Biophysics*, 52(Volume 52, 2023):113–138, May 2023. ISSN 1936-122X, 1936-1238. doi: 10.1146/annurev-biophys-062722-093258. URL https://www.annualreviews.org/content/journals/10.1146/annurev-biophys-062722-093258. Publisher: Annual Reviews.

Nicolas Chopin, Omiros Papaspiliopoulos, et al. *An introduction to Sequential Monte Carlo*, volume 4. Springer, 2020.

Arnak S Dalalyan. 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.

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

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(87)91197-X. URL https://www.sciencedirect.com/science/article/pii/037026938791197X.

Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 793–797. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/dwivedi18a.html.

R. Fletcher. *Practical Methods of Optimization*. Wiley, 1 edition, May 2000. ISBN 978-0-471-91547-8 978-1-118-72320-3. doi: 10.1002/9781118723203. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781118723203.

Charles J Geyer. Markov chain monte carlo maximum likelihood. 1991.

Louis Grenioux, Alain Oliviero Durmus, Eric Moulines, and Marylou Gabrié. On sampling with approximate transport maps. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11698–11733. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/grenioux23a.html.

Louis Grenioux, Maxence Noble, Marylou Gabrié, and Alain Oliviero Durmus. Stochastic localization via iterative posterior sampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16337–16376. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/grenioux24a.html.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Ye He, Kevin Rojas, and Molei Tao. Zeroth-order sampling methods for non-log-concave distributions: Alleviating metastability by denoising diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=X3Aljulsw5.

Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014. URL http://jmlr.org/papers/v15/hoffman14a.html.

Xunpeng Huang, Hanze Dong, Yifan HAO, Yian Ma, and Tong Zhang. Reverse diffusion Monte Carlo. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=kIPEyMSdFV.

Xunpeng Huang, Difan Zou, Hanze Dong, Yi-An Ma, and Tong Zhang. Faster sampling without isoperimetry via diffusion-based Monte Carlo. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 2438–2493. PMLR, 30 Jun–03 Jul 2024b. URL https://proceedings.mlr.press/v247/huang24a.html.

Koji Hukushima and Koji Nemoto. Exchange monte carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996. doi: 10.1143/JPSJ.65.1604. URL https://doi.org/10.1143/JPSJ.65.1604.

Ghassen Jerfel, Serena Wang, Clara Wong-Fannjiang, Katherine A. Heller, Yian Ma, and Michael I. Jordan. Variational refinement for importance sampling using the forward Kullback-Leibler divergence. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 1819–1829. PMLR, December 2021. URL https://proceedings.mlr.press/v161/jerfel21a.html. ISSN: 2640-3498.

William L. Jorgensen. The Many Roles of Computation in Drug Discovery. *Science*, 303(5665): 1813–1818, March 2004. doi: 10.1126/science.1096361. URL https://www.science.org/doi/10.1126/science.1096361. Publisher: American Association for the Advancement of Science.

Tony Lelièvre, Gabriel Stoltz, and Mathias Rousset. *Free energy computations: a mathematical perspective*. Imperial College Press, London ; Hackensack, N.J, 2010. ISBN 978-1-84816-247-1. OCLC: ocn244765923.

Tiancheng Li, Shudong Sun, Tariq Pervez Sattar, and Juan Manuel Corchado. Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches. *Expert Systems with applications*, 41(8):3944–3954, 2014.

David JC MacKay. Local minima, symmetry-breaking, and model pruning in variational free energy minimization. *Inference Group, Cavendish Laboratory, Cambridge, UK*, 2001.

Bálint Máté and François Fleuret. Learning interpolations between Boltzmann densities. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=TH6YrEcbth.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 06 1953. ISSN 0021-9606. doi: 10.1063/1.1699114. URL https://doi.org/10.1063/1.1699114.

Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023.

Thomas Müller, Brian Mcwilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. Neural importance sampling. *ACM Trans. Graph.*, 38(5), October 2019. ISSN 0730-0301. doi: 10.1145/3341156. URL https://doi.org/10.1145/3341156.

Maxence Noble, Louis Grenioux, Marylou Gabrié, and Alain Oliviero Durmus. Learned reference-based diffusion sampling for multi-modal distributions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fmJUYgmMbL.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006. ISBN 978-0-387-30303-1. doi: 10.1007/978-0-387-40065-5. URL http://link.springer.com/10.1007/978-0-387-40065-5.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019. doi: 10.1126/science.aaw1147. URL https://www.science.org/doi/abs/10.1126/science.aaw1147.

P. M. Pardalos and G. Schnitger. Checking local optimality in constrained quadratic programming is NP-hard. *Operations Research Letters*, 7(1):33–35, February 1988. ISSN 0167-6377. doi: 10.1016/0167-6377(88)90049-1. URL https://www.sciencedirect.com/science/article/pii/0167637788900491.

Gabriel Peyré and Marco Cuturi. 2019. doi: 10.1561/2200000073.

Danilo Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1530–1538. PMLR, June 2015. URL https://proceedings.mlr.press/v37/rezende15.html. ISSN: 1938-7228.

Lorenz Richter, Julius Berner, and Guan-Horng Liu. 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.

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

Zhekun Shi, Longlin Yu, Tianyu Xie, and Cheng Zhang. Diffusion-PINN sampler, 2024. URL https://arxiv.org/abs/2410.15336.

Roman Soletskyi, Marylou Gabrié, and Bruno Loureiro. A theoretical perspective on mode collapse in variational inference, 2024. URL https://arxiv.org/abs/2410.13300.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *The Ninth International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Vincent Stimper, David Liu, Andrew Campbell, Vincent Berenz, Lukas Ryll, Bernhard Schölkopf, and José Miguel Hernández-Lobato. normflows: A pytorch package for normalizing flows. *Journal of Open Source Software*, 8(86):5361, 2023. doi: 10.21105/joss.05361. URL https://doi.org/10.21105/joss.05361.

Robert H. Swendsen and Jian-Sheng Wang. Replica monte carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57:2607–2609, Nov 1986. doi: 10.1103/PhysRevLett.57.2607. URL https://link.aps.org/doi/10.1103/PhysRevLett.57.2607.

Francisco Vargas, Will Sussman Grathwohl, and Arnaud Doucet. Denoising diffusion samplers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=8pvnfTAbu1f.

Francisco Vargas, Shreyas Padhy, Denis Blessing, and Nikolas Nüsken. Transport meets Variational Inference: Controlled Monte Carlo Diffusions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PP1rudnxiW.

Ulli Wolff. Critical slowing down. *Nuclear Physics B - Proceedings Supplements*, 17:93–102, September 1990. ISSN 0920-5632. doi: 10.1016/0920-5632(90)90224-I. URL https://www.sciencedirect.com/science/article/pii/092056329090224I.

Yuling Yao, Aki Vehtari, and Andrew Gelman. Stacking for Non-mixing Bayesian Computations: The Curse and Blessing of Multimodal Posteriors. *Journal of Machine Learning Research*, 23 (79):1–45, 2022. ISSN 1533-7928. URL http://jmlr.org/papers/v23/20-1426.html.

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

A DETAILS ON CONSIDERED SAMPLERS

In the following, we present a brief description of the samplers presented in Section 3, and used in our numeric evaluation. Throughout this section, we will consider an arbitrary test function $f : \mathbb{R}^d \to \mathbb{R}$, for which we would like to evaluate the expectation $\mathbb{E}_\pi[f] := \int_{\mathbb{R}^d} f(x)\mathrm{d}\pi(x)$, and will explain how to compute Monte Carlo estimations $\mathbb{E}_\pi[f]$ for each sampler. Note that in our numerical experiment, we will take $f = \mathbb{1}_{S_1}$. Below, p^{init} will refer to an easy-to-sample distribution, with tractable density and normalization constant (basically, a Gaussian distribution).

MALA sampler. This algorithm is an extension of the largely used Unadjusted Langevin Algorithm (ULA) (Roberts & Tweedie, 1996), which builds a Markov chain $(X^\ell)_{\ell=0}^L$ with $(L+1)$ states $(L \geq 1)$, defined by $X^0 \sim p^{\text{init}}$ and for any $\ell \in \{0, \ldots, L-1\}$,

$$X^{\ell+1} = X^\ell + \lambda \nabla \log \gamma(X^\ell) + \sqrt{2\lambda} Z^\ell \ , \ Z^\ell \sim \mathrm{N}(0, \mathrm{I}_d) \ , \tag{2}$$

which reduces to a time discretization of the Langevin diffusion that admits π as invariant distribution. Building upon this recursion, MALA (Roberts & Tweedie, 1996) addresses the time discretization bias by adding a Metropolis-Hastings (MH) acceptance/rejection step at each iteration ℓ based on the acceptance probability

$$\alpha(x, x') = \min\left(1, \frac{\gamma(x')\mathrm{N}(x; x' + \lambda\nabla\log\gamma(x'), 2\lambda\mathrm{I}_d)}{\gamma(x)\mathrm{N}(x'; x + \lambda\nabla\log\gamma(x), 2\lambda\mathrm{I}_d)}\right) \ . \tag{3}$$

In this case, the corresponding Monte Carlo estimation of $\mathbb{E}_\pi[f]$ will be given by $L^{-1}\sum_{\ell=1}^L f(X^\ell)$.

Importance Sampling. Given an arbitrary proposal probability distribution ν with tractable density q known up to a normalizing constant such that $\pi \ll \nu$, Importance Sampling (IS) relies on the following equality

$$\mathbb{E}_\pi[f] = \mathbb{E}_\nu\left[\frac{\mathrm{d}\pi}{\mathrm{d}\nu} f\right] = \frac{1}{\mathcal{Z}}\mathbb{E}_\nu\left[\frac{\gamma}{q} f\right] \ . \tag{4}$$

Hence, the IS Monte Carlo estimation of $\mathbb{E}_\pi[f]$ is given by the weighted mean $N^{-1}\sum_{n=1}^N w_n(X^n) f(X^n)$, where $(X^n)_{n=1}^N$ are samples from ν and $\{w_n(X^n)\}_{n=1}^N$ are self-normalized importance weights defined by

$$w_n(X^n) = \frac{\tilde{w}_n(X^n)}{\sum_{m=1}^N \tilde{w}_m(X^m)} \ , \ \tilde{w}_n(X^n) = \frac{\gamma(X^n)}{q(X^n)} \ . \tag{5}$$

In the special case where the proposal is defined by a normalizing flow $\mathrm{T}_\theta : \mathbb{R}^d \to \mathbb{R}^d$ with base distribution p^{init}, the corresponding parametric density is given by

$$q_\theta(x) = p^{\text{init}}(\mathrm{T}_\theta^{-1}(x)) \left| J_{\mathrm{T}_\theta^{-1}}(x) \right| \ . \tag{6}$$

where $J_{\mathrm{T}_\theta^{-1}}$ denotes the Jacobian of T_θ^{-1}.

Variational Inference. Given a family of parametric distributions $\{\nu_\theta\}_{\theta\in\Theta}$ such that for any $\theta \in \Theta$, (i) ν_θ admits a tractable density q^θ known up to a normalizing constant and (ii) ν_θ is easy to sample from, VI aims to solve the following optimization problem

$$\theta^\star \in \arg\min_{\theta\in\Theta} \mathrm{D}(\nu_\theta|\pi) \ , \tag{7}$$

where D is a divergence[2] operating on the space of probability distributions. In most cases, D is defined as the reverse Kullback-Leibler divergence, i.e., we have $\mathrm{D}(\nu_\theta|\pi) = \mathrm{KL}(\nu_\theta|\pi) = \int_{\mathbb{R}^d} \log(q_\theta(x)/\gamma(x))\mathrm{d}\nu_\theta(x)$, where the equality holds up to constants independent from θ. In practice, this objective function is minimized by performing Stochastic Gradient Descent (SGD), based on samples from the variational distribution being trained. Assuming that $\hat{\theta}$ is the output of this numerical procedure, the corresponding VI Monte Carlo estimation of $\mathbb{E}_\pi[f]$ is given by $N^{-1}\sum_{n=1}^N f(X_{\hat{\theta}}^n)$, where $(X_{\hat{\theta}}^n)_{n=1}^N$ are samples from $\nu_{\hat{\theta}}$.

[2]Commonly, a divergence D verifies $\mathrm{D}(\nu \mid \pi) \geq 0$ and $\mathrm{D}(\nu \mid \pi) = 0$ if and only if $\nu = \pi$.

Annealed sampling. Given $K \geq 1$ the number of intermediate steps, the standard geometric density interpolation defines a sequence of target densities $\{p_k\}_{k=0}^K$ up to a normalizing constant by $p_k \propto (p^{\text{init}})^{1-\beta_k}\gamma^{\beta_k}$, where $k \in \{0, \dots, K\}$ and $\{\beta\}_{k=0}^K$ is an increasing sequence with $\beta_0 = 0$ and $\beta_K = 1$. In the literature, $\{\beta\}_{k=0}^K$ is often referred to as the tempering schedule.

Sequential Monte Carlo. Given $N \geq 1$, SMC generates N chains of size $K+1$, referred to as particles, such that the k-th elements of all chains are expected to be samples from p_k. To do so, SMC relies *locally* on MALA steps performed independently on each chain and *globally* on resampling steps based on accumulated importance weights, as defined next.

Let $k \in \{0, \dots, K-1\}$, $(\tilde{w}_k^n)_{n=1}^N$ be unnormalized importance weights accumulated on each chain up to step k ($\tilde{w}_k^n = 1/N$ if $k = 0$), and $(x_k^n)_{n=1}^N$ be samples from p_k obtained independently as the last samples of N parallel MALA runs (or exactly obtained if $k = 0$). Then, the SMC importance weights designed to sample from p_{k+1}, denoted by $(\tilde{w}_{k+1}^n)_{n=1}^N$, and their renormalized versions, denoted by $(w_{k+1}^n)_{n=1}^N$, are defined as

$$\tilde{w}_{k+1}^n = \frac{p_{k+1}(x_k^n)}{p_k(x_k^n)}\tilde{w}_k^n, \quad w_{k+1}^n = \frac{\tilde{w}_{k+1}^n}{\sum_{j=1}^N \tilde{w}_{k+1}^j} \ . \tag{8}$$

The SMC resampling step then consists in obtaining N new samples $(y_{k+1}^n)_{n=1}^N$ by random selection among the original samples $(x_k^n)_{n=1}^N$ via a multinomial distribution (with replacement) defined by the normalized weights $(w_{k+1}^n)_{n=1}^N$. These novel samples will then be used as starting points for running MALA on the target p_{k+1}. In our implementation, we follow the recommendation of Chopin et al. (2020) and build the sequence $\{\beta_k\}_{k=0}^K$ on-the-fly. This is done by solving, at each step k, the following optimization problem

$$\delta_k^\star = \min\{\delta \in [0, 1-\beta_k] : \text{ESS}(\delta) = \text{ESS}_{\text{min}}\} \ , \tag{9}$$

$$\text{where } \text{ESS}(\delta) = \left[\left(\sum_{n=1}^N \tilde{w}_{\delta,k}^n\right)^2 / \sum_{n=1}^N (\tilde{w}_{\delta,k}^n)^2\right] \quad \tilde{w}_{\delta,k}^n = \frac{(p^{\text{init}})^{1-(\beta_k+\delta)}(x_k^n)\gamma^{\beta_k+\delta}(x_k^n)}{p_k(x_k^n)} \ , \tag{10}$$

and $\text{ESS}_{\text{min}} = \alpha N$, with $\alpha \in (0, 1)$ being a user-defined threshold. Then, the schedule update is given by $\beta_{k+1} = \beta_k + \delta_k^\star$. By doing so, this procedure enables to prevent the degeneracy of the particles at step $k+1$ (which occurs when the corresponding ESS is too low). In practice, we minimize this quantity using Brent's method. In the end, the SMC estimator of $\mathbb{E}_\pi[f]$ is given by the weighted mean $N^{-1}\sum_{n=1}^N w_K^n f(X_K^n)$, where $(X_K^n)_{n=1}^N$ are the last MALA samples obtained at step K.

Replica Exchange. The RE algorithm aims at sampling from the annealed distributions $\{p_k\}_{k=0}^K$ with $K+1$ parallel Markov chains, i.e., the k-th chain targets p_k. For the sake of pedagogy, we assume here that K is even.

At each level k, RE sampling is done via a local MCMC sampler, such as MALA, which is expected to have poor performance for large k (as p_k will be highly multi-modal). To alleviate this issue, every L MALA steps, RE randomly performs a *swapping* between Markov chains. More specifically, each consecutive pair $(k, k+1)$ of Markov chains (i.e., either $(k, k+1) \in \{(0,1), \dots, (K-2, K-1)\}$ or $(k, k+1) \in \{(1,2), \dots, (K-1, K)\}$), with respective current states X_k and X_{k+1}, is swapped with probability

$$p_{k,k+1} = \min\left(1, \frac{p_k(X_{k+1})p_{k+1}(X_k)}{p_k(X_k)p_{k+1}(X_{k+1})}\right) \ . \tag{11}$$

After running the Markov chains for N steps, we obtain a Monte Carlo estimation of $\mathbb{E}_\pi[f]$ by considering $N^{-1}\sum_{n=1}^N f(X_K^n)$, where $(X_K^n)_{n=1}^N$ are the visited states of the K-th chain (corresponding to the target distribution).

SLIPS. Grenioux et al. (2024) proposed a diffusion-based sampling method extending the stochastic localization framework from Montanari (2023), which can be seen as an analog of denoising diffusion models. More precisely, given a horizon time $T > 0$, they consider a stochastic process $(Y_t)_{t\in[0,T]}$, called observation process, defined by

$$Y_t = \alpha(t)X + \sigma W_t \ , \ X \sim \pi \ , \tag{12}$$

where $\sigma \in (0, \infty)$, $\alpha : [0, T) \to [0, \infty)$ is an increasing function of time verifying $\alpha(0) = 0$ and $\alpha(t)/\sqrt{t} \to \infty$ as $t \to T$, and $(W_t)_{t\geq 0}$ is a d-dimensional Brownian motion. A particular choice of interest is given by $\alpha(t) = \sqrt{t/(1-t)}$, referred to as $\mathrm{Geom}(1, 1)$, with $T = 1$.

By definition of α, this observation process *localizes* on the target distribution as time increases since $Y_t/\alpha(t) \to X$ as $t \to T$. Hence, if we were able to simulate the stochastic process $(Y_t)_{t\in[0,T]}$ up to time T (which is infeasible a priori since we do not have access to ground truth samples), we would then obtain approximate samples from π. To do so, the authors propose to simulate the Markov process induced by the following SDE, which is proved to have the same marginals as the observation process, denoted by $(p_t)_{t\in[0,T]}$,

$$\mathrm{d}Y_t = \dot{\alpha}(t)u_t(Y_t)\mathrm{d}t + \sigma \mathrm{d}B_t \ , Y_0 = 0 \ , \tag{13}$$

where $(B_t)_{t\geq 0}$ is another d-dimensional Brownian motion and $u_t(y_t) = \mathbb{E}[X|Y_t = y_t]$ is the conditional expectation of the posterior distribution whose unnnormalized density is given by $q_t(x|y_t) : x \mapsto \mathrm{N}(y_t; \alpha(t)x, \sigma^2 t\mathrm{I}_d)\gamma(x)$. This drift being intractable, the authors propose to estimate it on-the-fly, i.e., while running a discretized version of the SDE (13), using MCMC methods to target the posterior q_t. They justify their method by highlighting the phenomenon *duality of log-concavity*, that is, while p_t gets increasingly multi-modal, q_t gets increasingly more log-concave and therefore, easier to sample from with standard MCMC techniques. Overall, the success of this method relies on tuning a critical hyperparameter $t_0 \in (0, T)$, which denotes the effective starting time of the generative SDE, such that both p_{t_0} and q_{t_0} are approximately log-concave. While some heuristics have been proposed by the authors to solve this issue, they may be not effective on arbitrary multi-modal distributions. For this algorithm, given a time-step t_1 close to T, the corresponding Monte Carlo estimation of $\mathbb{E}_\pi[f]$ is given by $N^{-1}\sum_{n=1}^N f(Y_{t_1}^n/\alpha(t_1))$, where $(Y_{t_1}^n)_{n=1}^N$ are obtained by running independently N times the approximated version of (13), up to time t_1.

DDS. This parameterized diffusion-based sampling method can be seen as the analog of the Variance-Preserving (VP) score-based generative model previously introduced by (Song et al., 2021), in the setting where ground truth samples are not available. More precisely, given a time horizon $T > 0$ and $\sigma \in (0, \infty)$, DDS aims to sample from a *denoising diffusion process* $(Y_t)_{t\in[0,T]}$ defined as the time-reversal of the Ornstein-Uhlenbeck process $(X_t)_{t\in[0,T]}$ induced by the following so-called *noising* SDE

$$\mathrm{d}X_t = -\frac{\beta(t)}{2}X_t\mathrm{d}t + \sigma\sqrt{\beta(t)}\mathrm{d}W_t \ , X_0 \sim \pi \ , \tag{14}$$

where $(W_t)_{t\geq 0}$ is a d-dimensional Brownian motion and $\beta : [0, T] \to (0, \infty)$ is a noising schedule chosen such that it holds approximately $X_T \sim \mathrm{N}(0, \sigma^2\mathrm{I}_d)$. In particular, it holds that $Y_t = X_{T-t}$ in the distribution sense, and, under mild assumptions on π and β, $(Y_t)_{t\in[0,T]}$ is solution to the following non-linear SDE

$$\mathrm{d}Y_t = \frac{\beta(T-t)}{2}Y_t\mathrm{d}t + \sigma^2\beta(T-t)\nabla \log p_{T-t}(Y_t)\mathrm{d}t + \sigma\sqrt{\beta(t)}\mathrm{d}B_t \ , Y_0 \sim \mathrm{N}(0, \sigma^2\mathrm{I}_d) \ , \tag{15}$$

where $(B_t)_{t\geq 0}$ is another d-dimensional Brownian motion and p_t denotes the density of the marginal distribution of X_t, which is intractable. Hence, in the same spirit as SLIPS, if we were able to simulate the stochastic process $(Y_t)_{t\in[0,T]}$ up to time T (which is infeasible a priori since the scores $\nabla \log p_t$ are intractable), we would then obtain approximate samples from π. To do so, Vargas et al. (2023) propose to learn the scores by solving the following Variational Inference problem, which does not require access to ground truth samples,

$$\theta^\star \in \arg\min_{\theta\in\Theta} \mathrm{KL}(\mathbb{P}^\theta|\mathbb{P}) \ , \tag{16}$$

where $\mathbb{P}$ is the path measure associated to the true reverse process, and for any $\theta \in \Theta$, $\mathbb{P}^\theta$ is the path measure associated to the neural SDE

$$\mathrm{d}Y_t = \frac{\beta(T-t)}{2} Y_t \mathrm{d}t + \sigma^2 \beta(T-t) s_\theta(T-t, Y_t)\mathrm{d}t + \sigma\sqrt{\beta(t)}\mathrm{d}B_t \ , Y_0 \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_d) \ , \quad (17)$$

where $s_\theta : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ is a neural network learnt such that $s_\theta(t,\cdot) \approx \nabla \log p_t$. To obtain a tractable objective function, Vargas et al. (2023) propose to parameterize s_θ via a linear term as $s_\theta(t,x) = -\frac{x}{\sigma^2} + u_\theta(t,x)$, where u_θ can be interpreted as a guidance. After time discretization, the optimization procedure to learn $\theta^\star$ amounts to perform VI on joint densities. Note that the reverse KL divergence may be replaced by the Log-Variance divergence for increased performance Richter et al. (2023). Assuming that $\hat{\theta}$ is the output of this numerical procedure, the corresponding DDS Monte Carlo estimation of $\mathbb{E}_\pi[f]$ is given by $N^{-1}\sum_{n=1}^N f(Y^n_{\hat{\theta},T})$ where $(Y^n_{\hat{\theta},T})_{n=1}^N$ are obtained by running independently N times the neural SDE (17) with control term $s_{\hat{\theta}}$, up to time T.

B Implementation details

Details on the multi-modal target distribution. In this paper, we consider the Gaussian mixture with density defined by

$$\gamma : x \in \mathbb{R}^d \mapsto \frac{2}{3}\overbrace{\mathrm{N}(x; -a\mathbf{1}_d, \Sigma_1)}^{q_1(x)} + \frac{1}{3}\underbrace{\mathrm{N}(x; a\mathbf{1}_d, \Sigma_2)}_{q_2(x)} \ , \quad (18)$$

where $\Sigma_1 \in \mathbb{R}^{d\times d}$ and $\Sigma_2 \in \mathbb{R}^{d\times d}$ are both diagonal matrices respectively defined by $(\Sigma_1)_{i,i} = \frac{i}{d}\sigma^2_{\max} + \frac{d-i}{d}\sigma^2_{\min}$ for $i \in \{1,\ldots,d\}$ and $(\Sigma_2)_{i,i} = (\Sigma_1)_{d-i,d-i}$ for $\sigma^2_{\max} = 0.2$ and $\sigma^2_{\min} = 0.01$. In our numerics, we considered the following range of hyperparameters: $a \in \{0.5, 2.875, 5.25, 7.625, 10\}$ and $d \in \{4, 8, 16, 32, 64\}$ (and the additional settings $d \in \{128, 256\}$ for SMC, RE and SLIPS). Below, we will consider the Gaussian approximation of π defined by $\mathrm{N}(\mathbf{m}_\pi, \Sigma_\pi)$, where $\mathbf{m}_\pi$ is the (tractable) mean of π and Σ_π is the diagonal positive matrix such that for any $i \in \{1,\ldots,d\}$, $\mathrm{Diag}(\Sigma_\pi)_i$ is the marginal variance of π on the i-th axis (which is also tractable).

Setting of MALA. To obtain one Monte Carlo estimation, we consider 32 independent chains randomly initialized in the strongest mode of π (i.e., in S_1). While running the Markov chains, we automatically adapt the step-size λ (initialized at 10^{-4}) by targeting a specific MH acceptance ratio fixed to 0.75. For each chain, we first run 4096 warm-up steps, and then run MALA for 8192 steps to obtain as many samples. This procedure produces 8192×32 samples which will be used to compute the mode weight estimation. We repeat this process 48 times to obtain aggregate results.

Setting of IS. To ensure an ideal setting for IS estimation, we set the proposal ν to be closely adapted to π, while avoiding giving access to the knowledge of the true mode proportions. This is done by minimizing the forward Kullback-Leibler divergence between the *equilibrated* version of π, defined by the mixture $x \mapsto \frac{1}{2}q_1(x) + \frac{1}{2}q_2(x)$, and the chosen parametric family of distributions (namely, Gaussian distributions or normalizing flows). Note that sampling from the equilibrated mixture is highly similar to directly sampling from the target distribution with an MCMC procedure, where the location of the modes is given. Regarding the parametric family used:

- Gaussian distributions: we compute maximum likelihood estimators of the ideal parameters, based on 16384 samples of the equilibrated target;
- Normalizing flows (NFs): we restrict our choice of parametrization to the RealNVP architecture (Dinh et al., 2017), which is composed of 32 affine coupling blocks with scale and shifts parametrized as Multi-Layer Perceptrons (MLPs) with 3 hidden layers of size 64. The base of the NF is a standard centered Gaussian distribution at temperature 0.5. The NF is trained on a dataset of 16384 samples for 100 epochs with a batch size of 1024 samples. We follow the implementation from Stimper et al. (2023).

Once training the proposal is done, we compute for each setting 48 IS estimators of the mode weight, using 8192 particles each time.

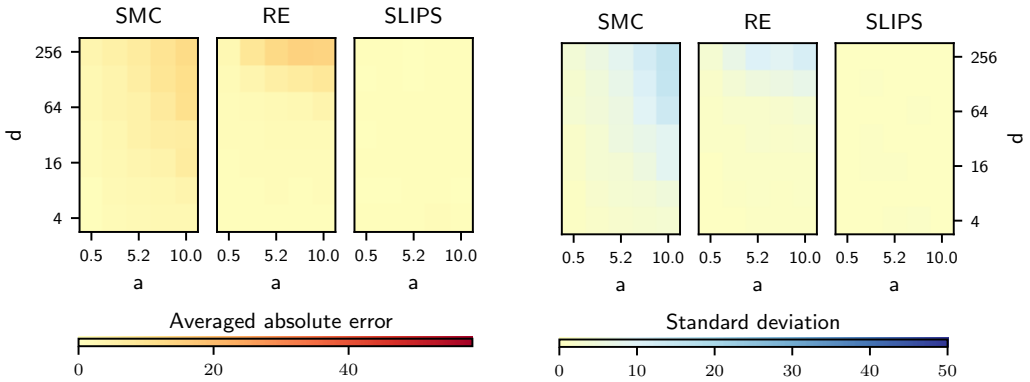

Figure 4: Same as Figure 2 for the best performing SMC, RE and SLIPS in higher-dimensions. Note that the color scale is the same as in Figure 2. **(Left):** Averaged absolute error of the estimation with respect to the true mode weight $w \approx 66.7\%$. **(Right):** Standard deviation of the estimation.

Setting of VI. We consider two different variational families :

• Gaussian distributions: we restrict our choice to distributions with diagonal covariances. The initial parameters of the VI procedure are taken as the exact mean of π and the diagonal of the exact covariance of π. In practice, we parameterize the log-variances instead of the variances;

• Normalizing flows (NFs): we restrict our choice of parametrization to the RealNVP architecture (Dinh et al., 2017), which is composed of 32 affine coupling blocks with scale and shifts parametrized as MLPs with 3 hidden layers of size 64. The base of the NF is a standard centered Gaussian distribution. We follow the implementation from Stimper et al. (2023).

In both cases, we minimize the reverse Kullback-Leibler divergence for 2048 SGD iterations with a batch size of 2048. Once the optimization is done, we sample 8192 from the obtained optimal parametric distribution to compute the mode weight estimation (this is done 48 times).

Setting of SMC. The base distribution p^{init} is given by $\mathrm{N}(\mathbf{m}_\pi, \Sigma_\pi)$. We set the maximum number of intermediate steps as 512 and $\mathrm{ESS}_{\text{min}} = N/2$ (i.e., $\alpha = 1/2$) as per Chopin et al. (2020). Locally, we perform MALA sampling, with adaptive step-size (initialized at 10^{-2}), for 96 steps. To ensure efficiency over the sequential procedure, we initialize the step-size of the $(k+1)$-th sampler with the last value of the step-size from the k-th sampler. Overall, we run SMC 48 times, with $N = 8192$ particles each time to compute the mode weight.

Setting of RE. The base distribution p^{init} is given by $\mathrm{N}(\mathbf{m}_\pi, \Sigma_\pi)$. We consider $K = 64$ intermediate steps and the schedule $\{\beta_k\}_{k=0}^K$ to be logarithmically spaced between $\varepsilon = 10^{-5}$ and 1, i.e., $\beta_k = 1 - \varepsilon^{(k/K)}$ for $k \in \{0, \dots, K\}$. In our experiments, we observed that this particular choice of schedule (unlike the linear schedule) has benefits to prevent the issue of mode switching in geometric density interpolations (Máté & Fleuret, 2023). Such as in SMC, we locally perform MALA sampling, with adaptive step-size (initialized at 10^{-2}). In particular, we enable the swapping every 8 MCMC steps. Overall, we perform 16384 warm-up steps at each annealing level and then 32768 steps to generate the samples, for 16 parallel instances of the algorithm. In the end, we perform a thinning by only selecting one sample every 8 steps in the K-th chain (corresponding to the target distribution), and randomly sub-sample 8192 samples out of all the obtained samples to compute the mode weight estimation. We repeat this procedure 48 times.

Setting of SLIPS. We run the SLIPS algorithm with the same setting as in the original paper, see Grenioux et al. (2024), but we rather use 512 discretization timesteps. Moreover, the Gaussian initialization procedure of SLIPS at time t_0 is preserved, but with the additional information of the true mean of the target distribution. As in the original paper, we tune the t_0 hyperparameter by leveraging access to ground truth samples. Overall, to obtain the Monte Carlo estimations, we run SLIPS 48×8192 times.

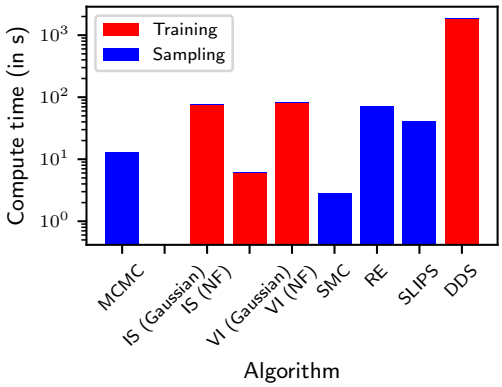

Figure 5: Average wall-clock computing time of each sampling algorithm. The results were averaged on 48×25 runs (48 Monte Carlo estimations for each of the 25 different mode separation/dimension settings).

Setting of DDS. We run the DDS sampler using the code from Berner et al. (2023), using 100 integration steps uniformly distributed between 0 and T. Following the guidelines from Vargas et al. (2023), we use a target-informed parametrization of the neural network u^θ. Moreover, the hyperparameter σ is set according to the recommendations of Noble et al. (2025)

$$\sigma^2 = \frac{1}{d}\left(m^\top m + \sum_{i=1}^{d} \Gamma_i^2\right) \tag{19}$$

where $m = \mathbb{E}_\pi[X] = a(2-w)\mathbf{1}_d$ and for $i \in \{1, \dots, d\}$,

$$\Gamma_i^2 = w\{(\Sigma_1)_{ii} + (a+m_i)^2\} + (1-w)\{(\Sigma_2)_{ii} + (a-m_i)^2\}\ . \tag{20}$$

Besides this, we train 16 times the DDS log-variance objective for 4096 gradient steps using a batch size of 1024 samples, and for each training, and compute, for each training, 3 Monte Carlo estimations of the mode weight based on 8192 generated samples (hence, overall 48 MC estimations).

Computational footprint. Figure 5 reports the average computing time of each algorithm. Note that, because of the on-the-fly discretization, the computational budget of SMC is stochastic. By way of indication, the budget of RE, SLIPS as well as the NF approaches would align with a 512 steps SMC algorithm.