A Practical Diffusion Path for Sampling

Omar Chehab [1] **Anna Korba** [1]

Abstract

Diffusion models are state-of-the-art methods in generative modeling when samples from a target probability distribution are available, and can be efficiently sampled, using score matching to estimate score vectors guiding a Langevin process. However, in the setting where samples from the target are not available, e.g. when this target's density is known up to a normalization constant, the score estimation task is challenging. Previous approaches rely on Monte Carlo estimators that are either computationally heavy to implement or sample-inefficient. In this work, we propose a computationally attractive alternative, relying on the so-called dilation path, that yields score vectors that are available in closed-form. This path interpolates between a Dirac and the target distribution using a convolution. We propose a simple implementation of Langevin dynamics guided by the dilation path, using adaptive step-sizes. We illustrate the results of our sampling method on a range of tasks, and shows it performs better than classical alternatives.

1. Introduction

Drawing samples from a target distribution is a key problem in statistics. In many settings, the target distribution is known up to a normalizing constant. This is often the case for pre-trained energy-based probabilistic models (Murphy, 2023, Chapter 24), whose parameters have already been estimated (Hyvärinen, 2005; Gutmann and Hyvärinen, 2012; Hinton, 2002; Gao et al., 2020). Another example comes from Bayesian statistics, where the posterior model over parameters given observed data is classically known only up to a normalizing constant (Wasserman, 2010, Chapter 11).

Classical methods for sampling from such target distributions, such as simulating a Langevin process with particles, are known to struggle when the target has many modes. Typically, the particles are first drawn to certain modes and then take exponential time to find all other modes (Bovier et al., 2000; 2004; 2005). Many successful, alternative methods rely on a path of distributions, chosen by the user to steer the sampling process to reach all the modes and hopefully converge faster (Neal, 2001; Geyer, 1991; Marinari and Parisi, 1992; Dai et al., 2020). An instance of such a method is Annealed Langevin Dynamics (Song and Ermon, 2019; 2020), that are simple to implement and are popular in Bayesian inference (Dai et al., 2020, Eq. 2.4), in global optimization (Geman and Hwang, 1986), and more recently in sampling from high-dimensional image distributions with many modes (Song and Ermon, 2019; 2020; Song et al., 2020). In the latter application, Annealed Langevin Dynamics have achieved state-of-the-art results by following a specific path of distributions, obtained by interpolating the multi-modal target and a standard Gaussian distribution with a convolution.

Yet, despite its promising geometry, this convolutional path of distributions is not readily usable for Annealed Langevin Dynamics, whose implementation requires the score vectors of the path which are not available in closed-form. A number of estimators for these score vectors have recently been developed when the target distribution is only known by its unnormalized density, yet these estimators can be computationally heavy (Huang et al., 2024a) or sample-inefficient (Huang et al., 2024b).

In this work, we introduce the dilation path, which is a limit case of the popular convolutional path in which the score vectors are available in closed-form. Our approach circumvents alternatives that require Monte Carlo simulation (Huang et al., 2024b; He et al., 2024; Grenioux et al., 2024; Saremi et al., 2024; Akhound-Sadegh et al., 2024) and is instead exceedingly simple to implement.

2. Background

Langevin dynamics The Unadjusted Langevin Algorithm (ULA) is a classical algorithm to draw samples from

[1] ENSAE, CREST, IP Paris, France. Correspondence to: Omar Chehab <emir.chehab@ensae.fr>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria. Copyright 2024 by the author(s).

a target distribution π. It is written as noisy gradient ascent

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + h_k \nabla \log \pi(\boldsymbol{x}_k) + \sqrt{2h_k}\boldsymbol{\epsilon}_k \tag{1}$$

with step sizes $h_k > 0$ and random noise $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. It can be viewed as a time-discretization of a Langevin diffusion (Borodin, 2017), where time is defined as $t = kh_k$ in the limit of null step sizes $h_k \to 0$. The rate of convergence is determined by constants (*e.g.* Log-Sobolev, Poincaré) that describe the geometry of the target distribution (Vempala and Wibisono, 2019): for multimodal distributions such as Gaussian mixtures, the constant degrades exponentially fast with the distance between modes (Holley and Stroock, 1987; Arnold et al., 2000), making convergence with given precision too slow to occur in a reasonable number of iterations. Yet, Langevin dynamics remain a popular choice for their computational simplicity: simulating Eq. 1 requires computing the score vector $\nabla \log \pi(\boldsymbol{x}_k)$ which does not depend on the target's normalizing constant. Hence, the Langevin sampler can be used to sample target distributions whose normalizing factor is unknown, and this property is unique among a broad class of samplers (Chen et al., 2023).

Annealed Langevin dynamics Many heuristics broadly known as annealing or tempering, consist in using the Langevin dynamics to sample from a path of distributions $(\mu_t)_{t\in\mathbb{R}_+}$ instead of the single target π, in hope that these intermediate distributions decompose the original sampling problem into easier tasks for Langevin dynamics. We specifically consider

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + h_k \nabla \log \mu_k(\boldsymbol{x}_k) + \sqrt{2h_k}\boldsymbol{\epsilon}_k, \tag{2}$$

where the target now moves with time. This process is known as Annealed Langevin Dynamics (Song and Ermon, 2019), and is sometimes combined with other sampling processes based on resampling (Dai et al., 2020, Eq. 2.4) or that directly simulate the path $(\mu_t)_{t\in\mathbb{R}_+}$ (Song et al., 2020, Appendix G).

Convolutional path Recently, a path obtained by taking the convolution of the target distribution π and an easier, proposal distribution ν

$$\mu_t(\boldsymbol{x}) = \frac{1}{\sqrt{1-\lambda_t}}\nu\left(\frac{\boldsymbol{x}}{\sqrt{1-\lambda_t}}\right) * \frac{1}{\sqrt{\lambda_t}}\pi\left(\frac{\boldsymbol{x}}{\sqrt{\lambda_t}}\right) \tag{3}$$

has produced state-of-the-art results in sampling from high-dimensional and multimodal distributions (Song and Ermon, 2019). Here, ν is typically a standard Gaussian and $\lambda : \mathbb{R}_+ \to [0, 1]$ in an increasing function called schedule (Chen, 2023): popular choices use exponential $\lambda_t = \min(1, e^{-2(T-t)})$ for some fixed $T \geq 0$, or linear $\lambda_t = \min(1, t)$ functions (Gao et al., 2023, Table 1). Note that the exponential schedule is initialized at $\lambda_0 = e^{-2T}$: choosing T to be big (resp. small) initializes the path of guiding distributions nearer to the proposal (resp. target). Recent work has empirically observed that this convolutional path may have a more favorable geometry for the Langevin sampler than another well-established path (Phillips et al., 2024), obtained by taking the geometric mean of the proposal and target distributions (Neal, 1998; Gelman and Meng, 1998a; Dai et al., 2020). However, using the convolutional path in practice requires computing the score vectors $\nabla \log \mu_t(\cdot)$ which has been at the center of recent work.

Computing the score with access to samples In machine learning literature, the target distribution is often accessed through samples $\boldsymbol{x}_\pi \sim \pi$ only. These can be interpolated with samples from the proposal distribution $\boldsymbol{x}_\nu \sim \nu$ to produce samples from the intermediate distributions μ_t (Albergo et al., 2023),

$$\boldsymbol{x}_t = \sqrt{1-\lambda_t}\boldsymbol{x}_\nu + \sqrt{\lambda_t}\boldsymbol{x}_\pi \ . \tag{4}$$

These samples can then be used to evaluate loss functions whose minimizers are estimators of the scores $\nabla \log \mu_t$ (Hyvärinen, 2005; Song and Ermon, 2019). Recent work provides theoretical guarantees that the estimation error deterioriates favorably, that is polynomially as opposed to exponentially, with the dimensionality of the data and separation between modes (Qin and Risteski, 2023).

Computing the score with access to an unnormalized density In classical statistics, it is instead assumed that the target distribution is accessed through its unnormalized density, not its samples. A novel way to compute the scores has been at the center of recent efforts to use the convolutional path in this setup (Huang et al., 2024b; He et al., 2024; Grenioux et al., 2024; Saremi et al., 2024; Akhound-Sadegh et al., 2024). Overwhelmingly, these works use an explicit Monte Carlo estimator of the score

$$\nabla \log \mu_t(\boldsymbol{x}) = \frac{e^{-(T-t)}}{1-e^{-2(T-t)}}\mathbb{E}_{\boldsymbol{y}\sim m_t}\left[\boldsymbol{y} - e^{T-t}\boldsymbol{x}\right], \tag{5}$$

$$m_t(\boldsymbol{y}|\boldsymbol{x}) \propto \pi(\boldsymbol{y}) \times \mathcal{N}(\boldsymbol{y}; e^{T-t}\boldsymbol{x}, \sqrt{e^{2(T-t)}-1}\boldsymbol{I}) \tag{6}$$

obtained by replacing the expectation with an average over finite samples. These samples are drawn from a blurred version of the target $m_t(\boldsymbol{y}|\boldsymbol{x})$, specifically the (normalized) product of the target and proposal distributions.

Using this estimator presents three important challenges:

1. The number of sampling procedures

   Each query of the score Eq. 5 at a certain $\boldsymbol{x}$ requires running a new sampling procedure Eq. 6. For example, each run of the algorithm Eq. 2 will query the score function at each iteration, and will require running as many sampling procedures.

2. The complexity of sampling procedures

   Standard sampling procedures for Eq. 6 are slow to converge. For example, He et al. (2024) use the rejection sampling algorithm whose computational cost is exponential in the dimension. Alternatively, the Unadjusted Langevin Algorithm Eq. 1 is guaranteed to be fast-converging when the sampled distribution Eq. 6 is log-concave, which is the case whenever the Gaussian distribution dominates the target. This is achieved by a small enough T in Eq. 5- 6, or equivalently by an initial schedule $\lambda_0 = e^{-2T}$ close enough to one so that the sampling process starts near the target (Huang et al., 2024b). Some works suppose that small enough T can be found as as a hyperparameter of the problem (Huang et al., 2024a; Grenioux et al., 2024). Using that value, Grenioux et al. (2024) estimate the score Eq. 5 over a window near the target $\lambda_t \in [e^{-2T}, 1]$. To estimate the score nearer the proposal, Huang et al. (2024a) use the distribution at e^{-2T} as the new target, and repeat. The computational complexity of such a procedure is prohitive: it scales exponentially in the number of windows (we verify this in Appendix A). Some sampling methods will altogether hide the difficulty of sampling a non log-concave distribution Eq. 6 by supposing access to an oracle (Lee et al., 2021; Chen et al., 2024).

3. The complexity of the estimator

   Even if we were able to efficiently sample Eq. 6, the estimator Eq. 5 introduces an estimation error in the sampling process that can scale exponentially with the dimensionality of data points (Huang et al., 2024b).

Importantly, these three challenges disappear when the score vectors $\nabla \log \mu_t(\cdot)$ are analytically computable. Finding a path that has both the favorable the geometry of the convolutional path and analytically computable score vectors is an unresolved problem.

3. The dilation path

We propose to use the limit of a Dirac proposal, anticipating that it will simplify the convolution which defines the path and consequently the score. We call the corresponding path a dilation path

$$\mu_t(\boldsymbol{x}) = \frac{1}{\sqrt{\lambda_t}} \pi \left(\frac{\boldsymbol{x}}{\sqrt{\lambda_t}} \right) \quad . \tag{7}$$

Notably, the scores are now analytically available

$$\nabla \log \mu_t(\boldsymbol{x}) = \frac{1}{\sqrt{\lambda_t}} \nabla \log \pi \left(\frac{\boldsymbol{x}}{\sqrt{\lambda_t}} \right) \quad . \tag{8}$$

We therefore recommend this path for practitioners and verify its simplicity in the experiments of section 4. We also note that considering a Dirac proposal has been known to simplify the analytical equation of a path from another problem known as Dynamic Optimal Transport or Schrödinger bridge. In that setup, the goal is to find a path that looks random (is close in Kullback-Leibler divergence to a Wiener process) but with fixed distributions at the start and end (proposal and target). Considering a Dirac proposal simplifies the equations of that path which is then given the name of Föllmer path (Ding et al., 2023; Huang et al., 2021; Jiao et al., 2021).

To better understand the geometry of the dilation path, we write the convolutional path between the proposal and target distributions assuming a fairly general parametric family and then consider the special case of a Dirac proposal. We recall the following in Appendix C

Proposition 1 (Gaussian mixture parametric family) *Suppose the proposal is a Gaussian $\nu := \mathcal{N}(\cdot, \mathbf{0}, \mathbf{\Sigma}_0)$ and the target is a mixture of M Gaussians with means $(\boldsymbol{\mu}_m)_{m \in [\![1,M]\!]}$, covariances $(\mathbf{\Sigma}_m)_{m \in [\![1,M]\!]}$, and positive weights $(w_m)_{m \in [\![1,M]\!]}$ that sum to one.*

The convolutional path Eq. 3 produces distributions π_t that conveniently remain in the Gaussian mixture parametric family. Their weights are constant $w_m(t) = w_m$ and their means and covariances interpolate additively the proposal's and target's, as $\boldsymbol{\mu}_m(t) = \sqrt{\lambda_t}\boldsymbol{\mu}_m$ and $\mathbf{\Sigma}_m(t) = (1 - \lambda_t)\mathbf{\Sigma}_0 + \lambda_t \mathbf{\Sigma}_m$.

In the case of a Dirac proposal, we have instead $\mathbf{\Sigma}_m(t) = \lambda_t \mathbf{\Sigma}_m$.

Along the interpolation between a Dirac and the target, the dilation path remains a mixture with constant weights, and only the means and covariances are updated. Preserving the mode weights along the path is a desirable feature for at least two reasons. First, it has been observed that the score vector used to simulate Eq. 2 is known to be rather blind to mode weights when the modes are far apart (Wenliang and Kanagawa, 2021). Second, it is reported that Langevin dynamics are slow-converging when initial and target weights differ, which is sometimes referred to as "mode switching" (Phillips et al., 2024).

Numerical implementation A naive implementation of the annealed Langevin sampler in Eq. 2 with the dilation path is numerically unstable: distributions p_{λ_t} when λ_t is close to zero have steep modes around which the gradient is numerically infinite. We verified in simple simulations that the particles diverge. To mitigate this effect, we use the effective numerical trick of an adaptive step size h to control the magnitude of the gradient term $h\nabla \log \pi_k(\boldsymbol{x})$ in Eq. 2.

Recent works have adapted the step size to time, so that $h_k \propto 1/\mathbb{E}[\|\nabla \log \pi_k(\boldsymbol{x}_k)\|^2]$, empirically observing that it decreases with the number of iterations (Song and Ermon, 2019, Section 4.3) (Song and Ermon, 2020, Section 3.3). Their motivation is to roughly equalize magnitudes of the gradient and noise terms in Eq. 2: $h_k\|\nabla \log \pi_k(\boldsymbol{x}_k)\| \approx \sqrt{2h_k}\|\epsilon_k\| \approx 1/\mathbb{E}[\|\nabla \log \pi_k(\boldsymbol{x}_k)\|]$. In practice, they use a proxy for $\mathbb{E}[\|\nabla \log \pi_k(\boldsymbol{x}_k)\|^2]$ which is not computed.

Instead, we use a finer adaptation, where the step size depends on time *as well as* the current position of a particle $h_k(\boldsymbol{x}_k) \propto 1/\|\nabla \log \pi_k(\boldsymbol{x}_k)\|$, or a bounded version in practice (Leroy et al., 2024, Eq. 3.1). Our motivation is to normalize the magnitude of the gradient term in Eq. 2. Note that adapting the step size in this way alters the stationary distribution of the sampling process. A corrective step can be used (Leroy et al., 2024, Eq. 2.14) but it involves a Hessian $\nabla^2 \log \pi(\cdot)$ which is expensive to compute in high dimensions and is outside the scope of this paper.

4. Experiments

In this section, we numerically verify convergence using the dilation path. We constrast it with another well-established path — obtained as the geometric mean of the proposal and target — whose scores are also available so that the computational budget for running Annealed Langevin Dynamics Eq. 2 is comparable.

Convergence metrics To measure the discrepancy between the distribution the particles p_k realizing the sampling process Eq. 2 and the target distribution π, we use (approximations of) a number of statistical divergences. Among them, we distinguish the Kernel Stein Discrepancy (KSD) (Liu et al., 2016; Chwialkowski et al., 2016; Gorham and Mackey, 2017), defined as

$$\mathrm{KSD}^2(p_k, \pi) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{x}')\sim p_k \otimes p_k}[K(\boldsymbol{x}, \boldsymbol{x}'; \nabla \log \pi, K')] \tag{9}$$

where K and K' are kernels defined in Appendix D. Importantly, the KSD can be approximated using what is available: samples from p_k and the score of the target π.

The other statistical divergences we use are abbreviated as (KL, revKL, MMD, OT) and are defined in Appendix D.1. Their approximations require access to samples from both the process p_k and the target π, which is not realistic as samples from the target are unavailable in practice. However, in our synthetic experiments, we are able to track these metrics.

We note many of these statistical divergences can be blind to mode coverage, which means that a sampling process can find few modes of the target π while ignoring other modes, and still produce low values of these metrics. In particular, this has been noted for the KSD (Korba et al., 2021; C. Benard, 2023) and the revKL (Verine et al., 2023). We therefore introduce a metric which we call the Multimodality Score (MMS), that specifically measures mode coverage. The MMS is defined as the root mean squared error between the actual and expected number of particles per mode.

Sampling a Gaussian mixture We follow the setups of Zhang et al. (2020) and Midgley et al. (2023) where the target distribution is a Gaussian mixture with 16 and 40 modes respectively. We use a standard Gaussian as the proposal distribution, except for the dilation path which has a Dirac proposal by design. Results are reported in Figure 1 and additional convergence diagnostics are reported in Appendix D.

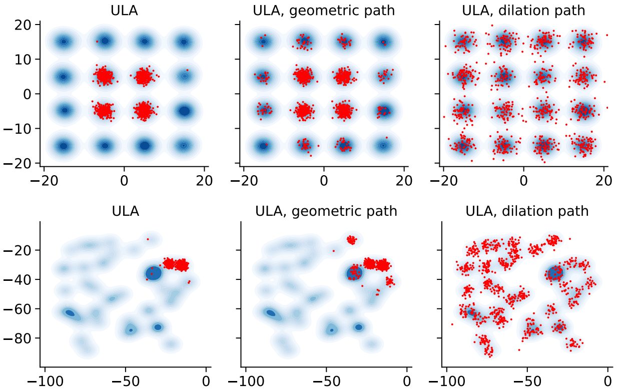

Figure 1: *Top.* 16-mode Gaussian mixture target and standard Gaussian proposal. *Bottom.* 40-mode Gaussian mixture target. The proposal is a standard Gaussian for ULA with/out the geometric path, and is a Dirac for ULA with the dilation path. The kernel density estimate of the target distribution is in blue; particles generated by the sampling process are in red. Simulations involved 1000 particles, 10000 iterations, a step size of 0.001, and a linear schedule.

These experiments highlight known benefits of Annealed Langevin Dynamics that follow a convolutional path, rather than an alternative, geometric path or no path at all. As expected in this setting, ULA with and without the geometric path is stuck for many iterations in modes that are closest to the initialization. One may be tempted to improve consider a proposal distribution with wider coverage so that more modes are already reached at initialization, but without knowing the locations and weights of the modes, the choice of "large enough" a Gaussian variance would be arbitrary. For example, some modes of the target could always be left in the tails of a proposal with larger variance. In contrast, ULA with the dilation path manages to recover all modes. This is reflected in most metrics in Figure 2.

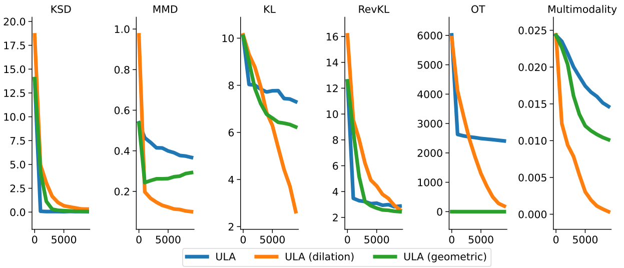

Figure 2: Convergence diagnostics for sampling the 40-mode Gaussian mixture target.

Sampling images Here, we use the dataset of images MNIST (Deng, 2012). Each image has a resolution of 28×28 pixels and can therefore be understood as a vector in a high-dimensional space $\mathbb{R}^{784}$ with entries in $[\![0, 255]\!]$. The target distribution is estimated from the MNIST dataset using a score-based diffusion model following Song and Ermon (2019). Results are reported in Figure 3.

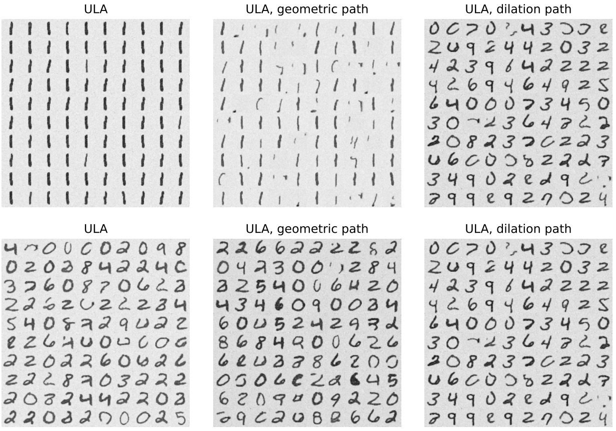

Figure 3: *Top*. The proposal is a standard Gaussian. *Bottom*. The proposal is a uniform distribution over the pixel domain. In both cases, the target distribution over images is estimated from the MNIST dataset using a score-based diffusion model following Song and Ermon (2019). Simulations involved 100 particles, 500 iterations, a step size of 0.001, and a linear schedule.

ULA with the dilation path consistently finds many modes, corresponding to different digits. The other sampling schemes, ULA with/out the geometric path, find modes in different proportions depending on the initialization: for example, when initialized near the origin at the top of Figure 3, they predominantly find the mode for digit "one", which is closest to the origin as we verify in Figure 4. However, with a uniform initialization, ULA with/out the geometric path finds more modes in more balanced proportions. A careful, quantitive study to understand how the mode proportions are affected by the initialization is left for future work.

5. Discussion

In this work, we introduced the dilation path, which a limit case of the popular convolutional path where the score vectors are available in closed-form. We show using this path for Annealed Langevin dynamics with a step size that is adaptive to both the time and position of a particle, yields an efficient and easy to implement sampler for multi-modal distributions. While we verified that this sampler recovers the locations of the modes (mode "coverage"), future work will be needed to verify if it correctly recovers their shapes as well (mode "fidelity").

We recall another path, similar in its derivation to the dilation path: it is also obtained by rescaling the target distribution (Ogata, 1990; 1996; Gelman and Meng, 1998b)

$$\mu_t(\boldsymbol{x}) = \sqrt{\lambda_t}\pi(\sqrt{\lambda_t}\boldsymbol{x}) \ . \tag{10}$$

except that the proposal distribution, defined in the limit of $\lambda_t \to 0$, now tends to a uniform distribution instead of a Dirac. This rescaled path will contract the modes from infinity inward, until they match the target modes when $\lambda_t = 1$ at a certain time. Because the mode locations remain far apart, Annealed Langevin dynamics could suffer from the mode separation (Bovier et al., 2000; 2004; 2005). This is in contrast to the dilation path, which starts with a Dirac and then expands the modes from the origin outward. Some works have also used this rescaled path for importance sampling, where the proposal is chosen in the family of distributions Eq. 10 for a certain value of λ (Sun et al., 2013), with the strong assumption that the target is Gaussian for tractability.

Acknowledgements Omar Chehab was supported by funding from the French ANR-20-CHIA-0016. The authors thank Nina Vesseron for sharing some code for Langevin sampling of images and Adrien Vacher for interesting conversations on the geometry of the dilation path.

References

K. P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL `http://probml.github.io/book2`.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

M. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(11):307–361, 2012.

G.E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8): 1771–1800, 2002.

R. Gao, E. Nijkamp, D.P. Kingma, Z. Xu, A.M. Dai, and Y. Nian Wu. Flow contrastive estimation of energy-based models. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7515–7525, 2020.

L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated, 2010. ISBN 1441923225.

A. Bovier, M. A. Eckhoff, V. Gayrard, and M. Klein. Metastability and low lying spectra in reversible markov chains. *Communications in Mathematical Physics*, 228: 219–255, 2000.

A. Bovier, M. A. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes i: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6:399–424, 2004.

A. Bovier, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes ii. precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7:69–99, 2005.

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

C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In E. M. Keramides, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163, 1991.

E. Marinari and G. Parisi. Simulated tempering: a new monte carlo scheme. *EPL*, 19:451–458, 1992.

C. Dai, J. Heng, P.E. Jacob, and N. Whiteley. An invitation to sequential monte carlo samplers. *Journal of the American Statistical Association*, 117:1587–1600, 2020.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Y. Song and S. Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12438–12448. Curran Associates, Inc., 2020.

S. Geman and C.-R. Hwang. Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24(5):1031–1043, 1986.

J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models, 2020.

X. Huang, D. Zou, H. Dong, Y. Ma, and T. Zhang. Faster sampling without isoperimetry via diffusion-based monte carlo. *ArXiv*, abs/2401.06325, 2024a.

X. Huang, H. Dong, Y. Hao, Y. Ma, and T. Zhang. Reverse diffusion monte carlo. In *International Conference on Learning Representations (ICLR)*, 2024b.

Y. He, K. Rojas, and M. Tao. Zeroth-order sampling methods for non-log-concave distributions: Alleviating metastability by denoising diffusion. *ArXiv*, abs/2402.17886, 2024.

L. Grenioux, M. Noble, M. Gabri'e, and A. Oliviero Durmus. Stochastic localization via iterative posterior sampling. *ArXiv*, abs/2402.10758, 2024.

S. Saremi, J. W. Park, and F. Bach. Chain of log-concave markov chains. In *International Conference on Learning Representations (ICLR)*, 2024.

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

A. N. Borodin. *Stochastic processes*. Springer, 2017.

S. Vempala and A. Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

R. Holley and D. W. Stroock. Logarithmic sobolev inequalities and stochastic ising models. *Journal of Statistical Physics*, 46:1159–1194, 1987.

A. Arnold, P. Markowich, and A. Unterreiter. On convex sobolev inequalities and the rate of convergence to equilibrium for fokker-planck type equations. *Communications in Partial Differential Equations*, 26, 05 2000.

Y. Chen, D. Zhengyu Huang, J. Huang, S. Reich, and A. M. Stuart. Gradient flows for sampling: Mean-field models, gaussian approximations and affine invariance, 2023.

T. Chen. On the importance of noise scheduling for diffusion models. *ArXiv*, abs/2301.10972, 2023.

Y. Gao, J. Huang, and Y. Jiao. Gaussian interpolation flows, 2023.

A. Phillips, H.-D. Dau, M. John Hutchinson, V. De Bortoli, G. Deligiannidis, and A. Doucet. Particle denoising diffusion sampler, 2024.

R.M. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 1998.

A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185, 1998a.

M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2023.

Y. Qin and A. Risteski. Fit like you sample: Sample-efficient generalized score matching from fast mixing diffusions, 2023.

Y. T. Lee, R. Shen, and K. Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pages 2993–3050. PMLR, 2021.

W. Chen, M. Zhang, B. Paige, J. Miguel Hernández-Lobato, and D. Barber. Diffusive gibbs sampling, 2024.

Z. Ding, Y. Jiao, X. Lu, Z. Yang, and C. Yuan. Sampling via föllmer flow, 2023.

J. Huang, Y. Jiao, L. Kang, X. Liao, J. Liu, and Y. Liu. Schrödinger-föllmer sampler: Sampling without ergodicity, 2021.

Y. Jiao, L. Kang, Y. Liu, and Y. Zhou. Convergence analysis of schrödinger-föllmer sampler without convexity, 2021.

L. K. Wenliang and H. Kanagawa. Blindness of score-based methods to isolated components and mixing proportions, 2021.

A. Leroy, B. Leimkuhler, J. Latz, and D. J. Higham. Adaptive stepsize algorithms for langevin dynamics, 2024.

Q. Liu, J. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pages 276–284. PMLR, 20–22 Jun 2016.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615. PMLR, 20–22 Jun 2016.

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, volume 70, pages 1292–1301. PMLR, 06–11 Aug 2017.

A. Korba, P.-C. Aubin-Frankowski, S. Majewski, and P. Ablin. Kernel stein discrepancy descent. In *International Conference on Machine Learning*, pages 5719–5730. PMLR, 2021.

S. Da Veiga C. Benard, B. Staber. Kernel stein discrepancy thinning: a theoretical perspective of pathologies and a practical fix with regularization. In *Neural Information Processing Systems (NeurIPS)*, 2023.

A. Verine, B. negrevergne, M. Sreenivas Pydi, and Y. Chevaleyre. Precision-recall divergence optimization for generative modeling with GANs and normalizing flows. In *Neural Information Processing Systems (NeurIPS)*, 2023.

R. Zhang, C. Li, J. Zhang, C. Chen, and A. Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. In *International Conference on Learning Representations (ICLR)*, 2020.

L. Illing Midgley, V. Stimper, G. N. C. Simm, B. Schölkopf, and J. Miguel Hernández-Lobato. Flow annealed importance sampling bootstrap. In *International Conference on Learning Representations (ICLR)*, 2023.

L. Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Y. Ogata. A monte carlo method for an objective bayesian procedure. *Annals of the Institute of Statistical Mathematics*, 42:403–433, 1990.

Y. Ogata. Evaluation of spatial bayesian models—two computational methods. *Journal of Statistical Planning and Inference*, 51(1):1–18, 1996. ISSN 0378-3758.

A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, pages 163–185, 1998b.

S. Sun, X. Li, H. Liu, K. Luo, and B. Gu. Fast statistical analysis of rare circuit failure events via scaled-sigma sampling for high-dimensional variation space. *2013 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 478–485, 2013.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research (JMLR)*, 13(25):723–773, 2012.

P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett,

J. Wilson, J.K. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, I. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

M. Cuturi, L. Meng-Papaxanthos, Y. Tian, C. Bunne, G. Davis, and O. Teboul. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

M. Waskom, O. Botvinnik, D. O'Kane, P. Hobson, S. Lukauskas, D. C Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, A. Miles, Y. Ram, T. Yarkoni, M. Lee Williams, C. Evans, C. Fitzgerald, B., C. Fonnesbeck, A. Lee, and A. Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), September 2017.

A. Computational complexity the recursive algorithm

We here discuss the computational complexity of the recurive algorithm of Huang et al. (2024a). We would like to approximate $\nabla \log \mu_k(\boldsymbol{x})$. We can do so using N_p particles sampled from Eq. 6 using ULA with N_i iterations. This will involve $N_p \times N_i$ evaluations of the target score $\nabla \log \pi(\cdot)$. Each of these scores can be approximated by repeating this procedure over N_s segments. The total computational complexity is therefore $(N_p \times N_i)^{N_s}$ queries of the final target score. This is exponential in the number of segments, which is computationally prohibitive.

B. Useful Lemma

Lemma 1 (Useful identities for a Gaussian density) *Interchangeability of mean and variance:* $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu};\boldsymbol{x},\boldsymbol{\Sigma})$.
Shift $\vec{\in}\mathbb{R}^d$, $\mathcal{N}(\boldsymbol{x} - \vec{;}\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu} + \vec{,}\boldsymbol{\Sigma})$.
Scaling $a \in \mathbb{R}$, $\mathcal{N}(\boldsymbol{x}/a;\boldsymbol{\mu},\boldsymbol{\Sigma}) = a\mathcal{N}(\boldsymbol{x};a\boldsymbol{\mu},a^2\boldsymbol{\Sigma})$.
Gradient $\nabla\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \times \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$.
Product $\mathcal{N}(x;\mu_1,\sigma_1^2) \times \mathcal{N}(x;\mu_2,\sigma_2^2) = \mathcal{N}(x;\frac{\mu_1\sigma_1^2+\mu_2\sigma_2^2}{\sigma_1^2+\sigma_2^2},\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2})$
Convolution $\mathcal{N}(x;\mu_1,\sigma_1^2) \times \mathcal{N}(x;\mu_2,\sigma_2^2) = \mathcal{N}(x;\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$.

C. Parametric families

We first prove Proposition C, restated here

Proposition (Gaussian mixture parametric family) *Suppose the proposal is a Gaussian* $\nu := \mathcal{N}(\cdot,\mathbf{0},\boldsymbol{\Sigma}_0)$ *and the target is a Gaussian mixture with* M *modes with means* $(\boldsymbol{\mu}_m)_{m\in[\![1,M]\!]}$, *covariances* $(\boldsymbol{\Sigma}_m)_{m\in[\![1,M]\!]}$, *and positive weights* $(w_m)_{m\in[\![1,M]\!]}$ *that sum to one.*

Then, distributions π_t *along the convolutional path remain in the Gaussian mixture parametric family. Their weights are constant* $w_m(t) = w_m$ *and their means and covariances interpolate additively the proposal's and target's, as* $\boldsymbol{\mu}_m(t) = \sqrt{t}\boldsymbol{\mu}_m$ *and* $\boldsymbol{\Sigma}_m(t)^2 = (1-t)\boldsymbol{\Sigma}_0 + t\boldsymbol{\Sigma}_m$.

Proof. Distributions along the convolutional path between the target to the proposal, are given by

$$\begin{aligned}
p_{\lambda_t}(\boldsymbol{x}) &= \frac{1}{\sqrt{1-\lambda_t}}\nu\Big(\frac{\boldsymbol{x}}{\sqrt{1-\lambda_t}}\Big) * \frac{1}{\sqrt{\lambda_t}}\pi\Big(\frac{\boldsymbol{x}}{\sqrt{\lambda_t}}\Big) \\
&= \frac{1}{\sqrt{1-\lambda_t}}\mathcal{N}\Big(\frac{\boldsymbol{x}}{\sqrt{1-\lambda_t}};\mathbf{0},\boldsymbol{\Sigma}_0\Big) * \sum_{i=1}^{M} w_m \frac{1}{\sqrt{\lambda_t}}\mathcal{N}\Big(\frac{\boldsymbol{x}}{\sqrt{\lambda_t}};\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m\Big) \\
&= \mathcal{N}(\boldsymbol{x};\mathbf{0},(1-\lambda_t)\boldsymbol{\Sigma}_0) * \sum_{i=1}^{M} w_m\mathcal{N}(\boldsymbol{x};\sqrt{\lambda_t}\boldsymbol{\mu}_m,\lambda_t\boldsymbol{\Sigma}_m) \\
&= \sum_{i=1}^{M} w_m\mathcal{N}(\boldsymbol{x};\mathbf{0},(1-\lambda_t)\boldsymbol{\Sigma}_0) * \mathcal{N}(\boldsymbol{x};\sqrt{\lambda_t}\boldsymbol{\mu}_m,\lambda_t\boldsymbol{\Sigma}_m) \\
&= \sum_{i=1}^{M} w_m\mathcal{N}(\boldsymbol{x};\sqrt{\lambda_t}\boldsymbol{\mu}_m,(1-\lambda_t)\boldsymbol{\Sigma}_0 + \lambda_t\boldsymbol{\Sigma}_m) \ .
\end{aligned}$$

They conveniently remain in the Gaussian mixture parametric family, and their parameters are

$$w_m(t) = w_m \quad \boldsymbol{\mu}_m(t) = \sqrt{\lambda_t}\boldsymbol{\mu}_m \quad \boldsymbol{\Sigma}_m(t) = (1-\lambda_t)\boldsymbol{\Sigma}_0 + \lambda_t\boldsymbol{\Sigma}_m \ . \tag{11}$$

If we consider the special case with the proposal covariance is $\boldsymbol{\Sigma}_0 = \epsilon\boldsymbol{I}$, then in the limit $\epsilon \to 0$ the proposal becomes a Dirac and we recover the dilation path with parameters

$$w_m(t) = w_m, \quad \boldsymbol{\mu}_m(t) = \sqrt{\lambda_t}\boldsymbol{\mu}_m, \quad \boldsymbol{\Sigma}_m(t) = \lambda_t\boldsymbol{\Sigma}_m \ . \tag{12}$$

□

D. Experiments

D.1. Statistical divergences used for evaluation

We next recall the definitions of the statistical divergences we use in our experiments, to measure the discrepancy between the distribution of particles p_k realizing the sampling process and the target distribution π.

- Kernel Stein Discrepancy (KSD) (Liu et al., 2016; Chwialkowski et al., 2016; Gorham and Mackey, 2017)

  Computing the KSD requires access to samples from p_k and to the density (more specifically, the score) of π.

  This statistical divergence is defined as

  $$\mathrm{KSD}^2(p_k, \pi) = \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{x}') \sim p_k \otimes p_k}[K(\boldsymbol{x}, \boldsymbol{x}'; \pi, K')] \tag{13}$$

  where K is a kernel whose computation requires the (unnormalized) target density π and another kernel K'

  $$K(\boldsymbol{x}, \boldsymbol{y}; \pi, K') = \nabla \log \pi(\boldsymbol{x})^T \nabla \log \pi(\boldsymbol{y}) K'(\boldsymbol{x}, \boldsymbol{y}) + \nabla \log \pi(\boldsymbol{x})^T \nabla_y K'(\boldsymbol{x}, \boldsymbol{y}) \tag{14}$$

  $$+ \nabla_x K'(\boldsymbol{x}, \boldsymbol{y})^T \nabla \log \pi(\boldsymbol{y}) + \nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{y}} K'(\boldsymbol{x}, \boldsymbol{y}) \ . \tag{15}$$

  chosen by the user. We use the recommended choice by Gorham and Mackey (2017), known as the Inverse Multiquadratic Kernel $K'(\boldsymbol{x}, \boldsymbol{y}) = (1 + \|\boldsymbol{x} - \boldsymbol{y}\|_2^2)^{-\beta}$ where $\beta \in [0, 1]$ and is here chosen as 0.5.

- Maximum Mean Discrepancy (MMD) (Gretton et al., 2012)

  Computing the MMD requires access to samples from both p_k and π.

  This statistical divergence is defined as

  $$\mathrm{MMD}^2(p_k, \pi) = 2\mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim p_k \otimes \pi}[K(\boldsymbol{x}, \boldsymbol{y})] - \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{x}') \sim p_k \otimes p_k}[K(\boldsymbol{x}, \boldsymbol{x}')] - \mathbb{E}_{(\boldsymbol{y}, \boldsymbol{y}') \sim \pi \otimes \pi}[K(\boldsymbol{y}, \boldsymbol{y}')] \tag{16}$$

  where K is a kernel chosen by the user. We use the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \exp\big(-\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 / 2h\big)$ with a bandwith $h = 1$.

- Kullback Leibler (KL) and reverse Kullback Leibler (revKL) divergences

  Computing the KL divergence usually requires access to samples from p_k and to the densities of both p_k and π. The reverse KL divergence is defined by switching the roles of p_k and π. In our experiments, we use an approximation implemented in the scipy library (Virtanen et al., 2020) that instead requires access to samples only, from both p_k and π.

  The KL divergence is defined as

  $$\mathrm{KL}(p_k, \pi) = \mathbb{E}_{\boldsymbol{x} \sim p_k}\left[\log \frac{p_k(\boldsymbol{x})}{\pi(\boldsymbol{x})}\right] \tag{17}$$

  and switching the roles of p_k and π yields the reverse KL divergence.

- 2-Wasserstein distance (OT)

  Computing the OT (which stands for Optimal Transport) requires access to samples from both p_k and π. In our experiments, we actually do not approximate directly the OT, but an upper bound using the Sinkhorn algorithm implemented in the ott library (Cuturi et al., 2022).

  The 2-Wasserstein distance is defined as

  $$W_2(p_k, \pi) = \inf_{c \in \mathcal{C}} E_{(\boldsymbol{x}, \boldsymbol{y}) \sim c}\left[\|\boldsymbol{x} - \boldsymbol{y}\|^2\right] \tag{18}$$

  where $\boldsymbol{x} \sim p_k$ and $\boldsymbol{y} \sim \pi$ and $\mathcal{C}$ is the set of joint distributions over such $(\boldsymbol{x}, \boldsymbol{y})$.

Last, we use the Multimodality Score (MMS) defined as the root mean squared error between the actual and expected number of particles per mode.

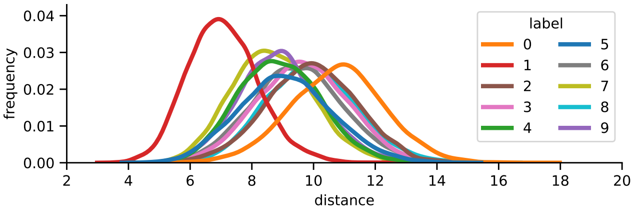

Figure 4: Distribution of the Euclidean distances of image vectors to the origin for the MNIST train dataset. We use the default kernel density estimate from the Seaborn python library (Waskom et al., 2017).

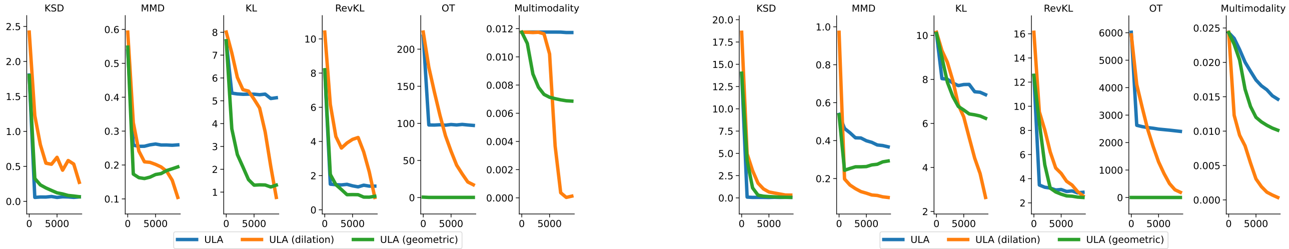

Figure 5: *Left*. Convergence diagnostics of the sixteen modes experiment. These divergences seem to be more sensitive to mode "fidelity" than mode "coverage", given that ULA seems to do better than ULA dilation. *Right*. Convergence diagnostics of the fourty modes experiment. The MMD and KL divergences seem to be sensitive to mode "coverage" where ULA dilation does best visually.