
NORMALIZING FLOW SAMPLING WITH LANGEVIN DYNAMICS IN THE LATENT SPACE

Florentin Coeurdoux, Nicolas Dobigeon*

IRIT/INP-ENSEEIHT
University of Toulouse
Toulouse, France

{Florentin.Coeurdoux, Nicolas.Dobigeon}@irit.fr

Pierre Chainais†

CNRS, Centrale Lille
University of Lille
Lille, France

Pierre.Chainais@centralelille.fr

ABSTRACT

Normalizing flows (NF) use a continuous generator to map a simple latent (e.g. Gaussian) distribution, towards an empirical target distribution associated with a training data set. Once trained by minimizing a variational objective, the learnt map provides an approximate generative model of the target distribution. Since standard NF implement differentiable maps, they may suffer from pathological behaviors when targeting complex distributions. For instance, such problems may appear for distributions on multi-component topologies or characterized by multiple modes with high probability regions separated by very unlikely areas. A typical symptom is the explosion of the Jacobian norm of the transformation in very low probability areas. This paper proposes to overcome this issue thanks to a new Markov chain Monte Carlo algorithm to sample from the target distribution in the latent domain before transporting it back to the target domain. The approach relies on a Metropolis adjusted Langevin algorithm (MALA) whose dynamics explicitly exploits the Jacobian of the transformation. Contrary to alternative approaches, the proposed strategy preserves the tractability of the likelihood and it does not require a specific training. Notably, it can be straightforwardly used with any pre-trained NF network, regardless of the architecture. Experiments conducted on synthetic and high-dimensional real data sets illustrate the efficiency of the method.

Keywords Normalizing flows, generative models, Monte Carlo sampling, Metropolis adjusted Langevin algorithm.

1 Introduction

Normalizing flows (NFs) are known to be a very efficient generative model to approximate probability distributions in an unsupervised setting. For example, *Glow* [Kingma and Dhariwal, 2018] is able to generate very realistic human faces, competing with state-of-the-art algorithms of variational inference [Papamakarios et al., 2021]. Despite some early theoretical results about their stability [Nalisnick et al., 2018] or their approximation and asymptotic properties [Behrmann et al., 2019], their training remains challenging in the most general cases. Their capacity is limited by intrinsic architectural constraints, resulting in a variational mismatch between the target distribution and the actually learnt distribution. In particular Cornish et al. [2020] pointed out the capital issue of target distributions with disconnected support featuring several components. Since NFs provide a continuous differentiable change of variable, they are not able to deal with such distributions when using a monomodal (e.g., Gaussian) latent distribution. Even targeting multimodal distributions featuring high probability regions separated by very unlikely areas remains problematic. The trained NF is a continuous differentiable transformation so that the transport of latent samples to the target space may overcharge low probability areas with (undesired) samples. These out-of-distribution samples will correspond to smooth transitions between different modes, which leads to out-of-distribution samples, as discussed by Cornish et al. [2020].

*This work was supported by the Artificial Natural Intelligence Toulouse Institute (ANITI, ANR-19-PI3A-0004).

†This work was supported by the AI Sherlock Chair (ANR-20-CHIA-0031-01), the ULNE national future investment programme (ANR-16-IDEX-0004) and the Hauts-de-France Region.

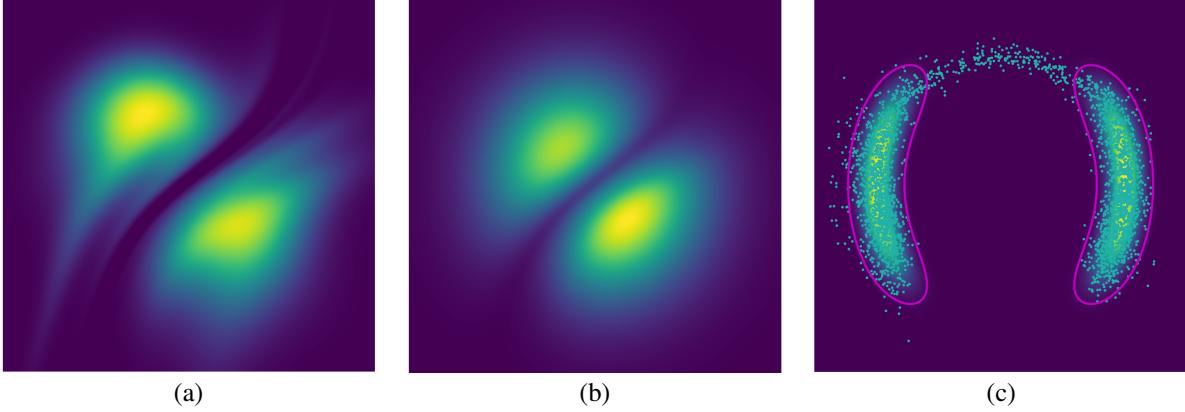


Figure 1: (a) Target distribution p_Z in the latent space; (b) Learnt latent distribution in the latent space \hat{q}_Z ; (c) Output of naive sampling, i.e., $\{x^{(n)}\}_{n=1}^N$ with $x^{(n)} = f(z^{(n)})$ and $z \sim q_Z$.

Fig. 1 illustrates this behavior on a archetypal bimodal 2D two-moon distribution using a latent Gaussian distribution. The NF is first trained on a large set of samples from the true target distribution. Fig. 1(a) shows the likelihood of the inverse transformation of the target distribution back to the latent space. It appears that the NF splits the Gaussian latent space into two sub-regions separated by an area of minimal likelihood. In the target domain, these minima correspond to the low probability area between the two modes of the target distribution, which is somewhat expected.

Fig. 1(c) shows the result of a naive sampling from the Gaussian model latent distribution to explore the target distribution thanks to the NF change of variable. Latent Gaussian independent samples are transformed by the NF to the target domain. The purple line represents the 97.5% level set. It appears that many samples through this naive sampling procedure are out-of-distribution in the low probability area between the two moons, see the top of the plot. They correspond to samples of the latent distribution around the low-likelihood area highlighted on Fig. 1(a). Note that this illustrative example and in particular Fig. 1(b) will be more deeply discussed in the contributive Sections 3 to 5 in the light of the findings reported along the paper.

The observations made above illustrate a behavior that is structural. NFs are diffeomorphisms that preserve the topological structure of the support of the latent distribution. If the information about the structure of the target distribution is ignored, many out-of-distribution samples will be generated. This effect is reinforced by the fact that the NF is trained on a finite data set so that in practice there exist close to empty areas in low probability regions. In other words, since the latent distribution is usually a simple Gaussian unimodal distribution, there is a topological mismatch with the often much more complex target distribution [Cornish et al., 2020], in particular when it is multimodal.

A first contribution of this paper is a theoretical study of the impact of a topological mismatch between the latent distribution on the Jacobian of the NF transformation. We prove that the norm of the Jacobian of a sequence of differentiable mappings between a unimodal distribution and a distribution with disconnected support diverges to infinity (see Proposition 1). This observation suggests that one should consider the information brought by the Jacobian when sampling from the target distribution with a NF. We also draw a connection between NF optimality and Gaussian isoperimetric theory [Ledoux, 1996]. This connection permits to explain the split of the latent space with minimal perimeter corresponding to Voronoi cells [Milman and Neeman, 2022].

Capitalizing on this theoretical study, the second contribution of this paper is a new dedicated Markov chain Monte Carlo algorithm to sample efficiently according to the distribution targeted by a NF. The proposed sampling method builds on a Langevin dynamics formulated in the target domain and translated into the latent space, which is made possible thanks to the invertibility of the NF. Interestingly the resulting Langevin diffusion is defined on the Riemann manifold whose geometry is driven by the Jacobian of the NF. As a result, the proposed Markov chain Monte Carlo method is shown to avoid low probability regions and to produce significantly less out-of-distribution samples, even when the target distribution is multimodal. It is worth noting that the proposed method does not require a specific training procedure but can be implemented to sample from any pre-trained NF with any architecture.

The paper is organized as follows. Section 2 reports on related works. Section 3 recalls the main useful notions about normalizing flows. Section 4 studies the theoretical implications of a topological mismatch between the latent distribution and the target distribution. Section 5 introduces the proposed sampling method based on a Langevin dynamics in the latent space. In Section 6, numerical experiments illustrate the advantages of the proposed approach

by reporting performance results both for 2D toy distributions and in high dimensions on the usual Cifar-10, CelebA and LSUN data sets.

2 Related works

Geometry in neural networks – Geometry in neural networks as a tool to understand local generalization was first discussed by Bengio et al. [2013]. As a key feature, the Jacobian matrix controls how smoothly a function interpolates a surface from some input data. As an extension, Rifai et al. [2011] showed that the norm of the Jacobian acts as a regularizer of the deterministic autoencoder. Later Arvanitidis et al. [2018] were the first to establish the link between push forward generative models and surface modeling. In particular, they showed that the latent space could reveal a distorted view of the input space that can be characterized by a stochastic Riemannian metric governed by the local Jacobian.

Distribution with disconnected support – As highlighted by Cornish et al. [2020], when using ancestral sampling, the structure of the latent distribution should fit the unknown structure of the target distribution. To tackle this issue, several solutions have been proposed. These strategies include augmenting the space on which the model operates [Huang et al., 2020], continuously indexing the flow layers [Cornish et al., 2020], and including stochastic [Wu et al., 2020] or surjective layers [Nielsen et al., 2020]. However, these approaches sacrifice the bijectivity of the flow transformation. In most cases, this sacrifice has dramatic impacts: the model is no longer tractable, memory savings during training are no longer possible [Gomez et al., 2017], and the model is no longer a perfect encoder-decoder pair. Other works have promoted the use of multimodal latent distributions [Izmailov et al., 2020, Ardizzone et al., 2020, Hagemann and Neumayer, 2021]. Nevertheless, rather than capturing the inherent multimodal nature of the target distribution, their primary motivation is to perform a classification task or to solve inverse problems with flow-based models. Papamakarios et al. [2017] has shown that choosing a mixture of Gaussians as a latent distribution could lead to an improvement of the fidelity to multimodal distributions. Alternatively, Pires and Figueiredo [2020] has studied the learning of a mixture of generators. Using a mutual information term, they encourage each generator to focus on a different submanifold so that the mixture covers the whole support. More recently, Stumper et al. [2022] predicted latent importance weights and proposed a sub-sampling method to avoid the generation of the most irrelevant samples. However, all these methods require to implement elaborated learning strategies which handle several sensitive hyperparameters or impose specific neural architectures. On the contrary, as emphasized earlier, the present approach does not require a specific training strategy, is computationally efficient, and can be implemented to any pre-trained NF.

Sampling with normalizing flows – Recently NFs have been used to facilitate the sampling from distributions with non-trivial geometries by transforming them into distributions that are easier to handle. To solve the problem, samplers that combine Monte Carlo methods with NF have been proposed. On the one hand, flows have been used as reparametrization maps that improve the geometry of the target distribution before running local conventional samplers such as Hamiltonian Monte Carlo (HMC) [Hoffman et al., 2019, Noé et al., 2019]. On the other hand, the push-forward of the NF base distribution through the map has also been used as an independent proposal in importance sampling [Müller et al., 2019] and Metropolis-Hastings steps [Gabrié et al., 2022, Samsonov et al., 2022]. In this context, NFs are trained using the reverse Kullback-Leiber divergence so that the push-forward distribution approximates the target distribution. These approaches are particularly appealing when a closed-form expression of the target distribution is available explicitly. In contrast, this paper does not assume an explicit knowledge of the target distribution. The proposed approach aims at improving the sampling from a distribution learnt by a given NF trained beforehand.

3 Normalizing flows: preliminaries and problem statement

3.1 Learning a change of variables

NFs define a flexible class of deep generative models that seek to learn a change of variable between a reference Gaussian measure q_Z and a target measure p_X through an invertible transformation $f : \mathcal{Z} \rightarrow \mathcal{X}$ with $f \in \mathcal{F}$ where \mathcal{F} defines the class of NFs. Fig. 2 summarizes the usual training of NF that minimizes a discrepancy measure between the target measure p_X and the push-forwarded measure q_X defined as

$$q_X = f_{\sharp} q_Z \tag{1}$$

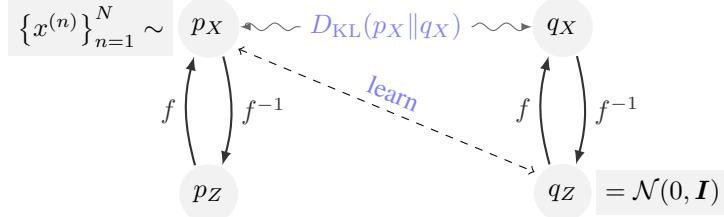


Figure 2: NF learns a mapping f from data points $\{x^{(n)}\}_{n=1}^N$ assumed to be drawn from p_X towards the latent Gaussian measure q_Z . The training consists in minimizing the KL divergence between p_X and $q_X = f_\sharp q_Z$. Once trained, the learnt map permits to go from q_Z to q_X , which is an approximation of the true target distribution p_X .

where f_\sharp stands for the associated push-forward operator. This discrepancy measure is generally chosen as the Kullback-Leibler (KL) divergence $D_{\text{KL}}(p_X \| q_X)$. Explicitly writing the change of variables

$$q_X(x) = q_Z(f^{-1}(x)) |J_{f^{-1}}(x)| \quad (2)$$

where $J_{f^{-1}}$ is the Jacobian matrix of f^{-1} , the training is thus formulated as the minimization problem

$$\min_{f \in \mathcal{F}} \mathbb{E}_{p_X} [-\log q_Z(f^{-1}(x)) + \log |J_{f^{-1}}(x)|] \quad (3)$$

Note that the term $\log p_X(x)$ does not appear in the objective function since this latter does not depend on f . In the present work, the class \mathcal{F} of admissible transformations is chosen as the structures composed of coupling layers (Papamakarios et al. [2021], Dinh et al. [2016], Kingma and Dhariwal [2018]) ensuring the Jacobian matrix of f to be lower triangular with positive diagonal entries. Because of this triangular structure, the Jacobian J_f and the inverse of the map f^{-1} are available explicitly. In particular the Jacobian determinant $|J_f(z)|$ evaluated at $z \in \mathcal{Z}$ measures the dilation, the change of volume of a small neighborhood around z induced by f , i.e., the ratio between the volumes of the corresponding neighborhoods of x and z .

In practice, the target measure p_X is available only though observed samples $\{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$. Adopting a sample-average approximation, the objective function in (3) is replaced by its Monte Carlo estimate. For this fixed set of samples per data batch, the NF training is formulated as

$$\hat{f} \in \min_{f \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^N \left[-\log q_Z(f^{-1}(x^{(n)})) + \log |J_{f^{-1}}(x^{(n)})| \right]. \quad (4)$$

In the following, it will be worth keeping in mind that the obtained solution \hat{f} is only an approximation of the exact transport map due to two reasons. First, the feasible set \mathcal{F} (i.e., the class of admissible NFs) is reduced to the set of continuous, differentiable and bijective functions. The existence of a transformation belonging to this class such that $D_{\text{KL}}(p_X \| q_X) = 0$ is not guaranteed. This remark holds even in the case of a parametric form of the target distribution p_X , i.e., when minimizing the real (non-empirical) objective function in (3). Second, even when such a transformation exists in \mathcal{F} , the solution \hat{f} recovered by (4) coincides with the minimizer of (3) in an asymptotic sense only, i.e., when $N \rightarrow \infty$. On top of that, only a proxy of \hat{f} can be computed because of the use of a stochastic optimization procedure (e.g., stochastic gradient descent), which may suffer from issues not discussed anyfurther here.

The main issues inherent to the NF training and identified above would still hold for more refined training procedures [Coeurduox et al., 2022], i.e., that would go beyond to the crude minimization problem (3). However, the work reported in this paper does not address the training of the NF. Instead, one will focus on the task which consists in generating samples from the learnt target measure. Thus one will assume that a NF has been already trained to learn a given change of variable. To make the sequel of this paper smoother to read, no distinction will be made between the sought transformation and its estimate, that will be denoted f in what follows.

3.2 A Gaussian latent space?

As noticed by Marzouk et al. [2016], learning the transformation f by variational inference can be reformulated with respect to (w.r.t.) the corresponding inverse map f^{-1} . Since the KL divergence is invariant to changes of variables, minimizing $D_{\text{KL}}(p_X \| q_X)$ is equivalent to minimizing $D_{\text{KL}}(p_Z \| q_Z)$ with $p_Z = f_\sharp^{-1} p_X$. The training procedure is thus formulated in the latent space instead of the target space. In other words, the NF aims at fitting the target measure

p_Z expressed in the latent space to the latent Gaussian measure q_Z . However, due to inescapable shortcomings similar to those highlighted above, the target measure p_Z in the latent space is only an approximation of the latent Gaussian measure q_Z . This mismatch can be easily observed in Fig. 1(a) where the depicted actual measure p_Z is clearly not Gaussian. This issue may be particularly critical when there is a topological mismatch between the respective supports of the target and latent distributions. This will be discussed in more details in Section 4.

3.3 Beyond conventional NF sampling

Once the NF has been trained, the standard method to sample from the learnt target distribution is straightforward. It consists in drawing a sample z_k from the latent Gaussian distribution q_Z and then applying the learnt transformation f to obtain a sample $x^{(n)} = f(z^{(n)})$. This method will be referred to as “naive sampling” in the sequel of this paper.

Unfortunately, as discussed in Section 3.2 (see also Fig. 1), the latent distribution q_Z is expected to be different from the actual target distribution p_Z expressed in the latent space. As suggested in the next section, this mismatch will be even more critical when it results from topological differences between the latent and target spaces. As a consequence the naive NF sampling is doomed to be suboptimal and to produce out-of-distribution samples, as illustrated in Fig. 1(c). In contrast, the approach proposed in Section 5 aims at devising an alternative sampling strategy that explicitly overcomes these shortcomings.

4 Implications of a topological mismatch

4.1 Topology preservation

The push-forward operator $f_\#$ learnt by an NF transports the mass allocated by q_Z in \mathcal{Z} to \mathcal{X} , thereby defining q_X by specifying where each elementary mass is transported. This imposes a global constraint on the operator f if the model distribution q_X is expected to match a given target measure p_X perfectly. Let $\text{supp}(q_Z) = \{z \in \mathcal{Z} : q_Z(z) > 0\}$ denote the support of q_Z . Then the push-forward operator $f_\#$ can yield $q_X = p_X$ only if

$$\text{supp}(p_X) = \overline{f(\text{supp}(q_Z))} \quad (5)$$

where \overline{B} is the closure of set B . The constraint (5) is especially onerous for NFs because of their bijectivity. The operators f and f^{-1} are continuous, and f is a homeomorphism. Consequently, for these models, q_Z and p_X are isomorphic, i.e., homeomorphic as topological spaces [Runde et al., 2005, Def. 3.3.10]. This means that $\text{supp}(q_Z)$ and $\text{supp}(p_X)$ must share exactly the same topological properties, in particular the number of connected components. This constraint may be unlikely satisfied when learning complex real-world distributions, leading to an insurmountable topological mismatch. In such cases, this finding has serious consequences on the operator f learnt and implemented by a NF. Indeed, the following proposition states that if the respective supports of the latent distribution q_Z and the target distribution p_X are not homeomorphic, then the norm of the Jacobian $|J_f|$ of f may become arbitrary large.

Here $\xrightarrow{\mathcal{D}}$ denotes weak convergence.

Proposition 1. *Let q_Z and p_X denote distributions defined on \mathbb{R}^d . Assume that $\text{supp}(q_Z) \neq \text{supp}(p_X)$. For any sequence of measurable, differentiable Lipschitz functions $f_t : \mathbb{R}^{d_Z} \rightarrow \mathbb{R}^{d_X}$, if $f_{t\#}q_Z \xrightarrow{\mathcal{D}} p_X$ when $t \rightarrow +\infty$, then*

$$\lim_{t \rightarrow \infty} \sup_{z \in \mathcal{Z}} (\|J_{f_t}(z)\|) = +\infty.$$

The proof is reported in Appendix A.

It is worth noting that training a generative model is generally conducted by minimizing a statistical divergence. For most used divergence measures, (e.g., KL and Jensen-Shannon divergences, Wasserstein distance), this minimization implies a weak convergence of the approximated distribution q_X towards the target distribution p_X [Arjovsky et al., 2017]. As a consequence, Proposition 1 states that, when training a NF to approximate p_X by $q_X = \lim_{t \rightarrow \infty} f_{t\#}q_Z$, the supremum of the Jacobian of the learnt mapping may become arbitrarily large in some regions. This result is in line with the experimental findings early discussed and visually illustrated by Fig. 1. Indeed, Fig. 1(b) depicts the heatmap of the log-likelihood

$$\log q_X(f(z)) = \log q_Z(z) - \log |J_f(z)| \quad (6)$$

given by (2) after training an NF. The impact of the term governed by the determinant of the Jacobian is clear. It highlights a boundary separating two distinct areas, each associated with a mode in the target distribution p_X . This result still holds when q_Z and q_X are defined on \mathbb{R}^{d_Z} and \mathbb{R}^{d_X} , respectively, with $d_Z \neq d_X$. This shortcoming is thus

also unavoidable when learning injective flow models [Kumar et al., 2017] and other push-forward models such as GANs [Goodfellow et al., 2020].

In practice, models are trained on a data set of finite size. In other words, the underlying target measure p_X is available only through the empirical measure $\frac{1}{N} \sum_{n=1}^N \delta_{x^{(n)}}$. During the training defined by (4), areas of low probability possibly characterizing a multi-modal target measure are likely interpreted as areas of null probability observed in the empirical measure. This directly results in the topological mismatch discussed above. Thus, even when targeting a distribution p_X defined over a connected support with regions of infinitesimal support, the learnt mapping is expected to be characterized by a Jacobian with exploding norm in these regions, see Fig 1. The following section suggests that these regions correspond to the frontiers between cells defining a partition of the latent space.

4.2 Partition of the latent space

A deterministic generative model can be interpreted as a surface model or a Gauss map, if the generator f is sufficiently smooth [Arvanitidis et al., 2018]. When targeting a multi-modal distribution, the learnt model implicitly partitions the Gaussian latent space into a set of disjoint subsets, each associated with a given mode. The Gaussian multi-bubble conjecture was formulated when looking for a way to partition the Gaussian space with the least-weighted perimeter. This conjecture was proven recently by Milman and Neeman [2022]. The result states that partitioning a Gaussian space of \mathbb{R}^d into m clusters of equal measures ($2 \leq m \leq d+1$) consists in recovering “simplicial clusters” defined as Voronoi cells of m equidistant points in \mathbb{R}^d . According to convex geometry principles, the boundary is the union of m convex cones, each of them contained in a distinct hyperplane that goes through the origin of \mathbb{R}^d . Recently, Issenhuth et al. [2022] leveraged on this finding to assess the optimality of the precision of GANs. They show that the precision of the generator vanishes when the number of components of the target distribution tends towards infinity.

Again, in practice, NFs models are trained on a data set with a finite number of samples. This results in a partitioning of the Gaussian latent space into cells separated by frontiers of arbitrarily large widths. Figure 1(b) allows a connection to be drawn between the statement by Milman and Neeman [2022] and the Proposition 1. Indeed, in this figure, the frontiers defining this partitioning are clearly identified as the areas with exploding Jacobian norm. As a consequence, exploring naively the Gaussian latent space to sample from the target distribution seems to be inappropriate. Because of these wide frontiers, large areas of the latent space are expected to be associated with potentially numerous out-of-distribution samples.

5 NF sampling in the latent space

5.1 Local exploration of the latent space

As explained in Section 3.3, naive NF sampling boils down to drawing a Gaussian variable before transformation by the learnt mapping f . This strategy is expected to produce out-of-distribution samples, due to the topological mismatch between q_X and p_X discussed in Section 4. The proposed alternative elaborates directly on the learnt target distribution q_X .

The starting point of our rational consists in expressing a Langevin diffusion in the target space. This Markov chain Monte Carlo (MCMC) algorithm would target the distribution q_X using only the derivative of its likelihood $\nabla_x \log q_X(x)$. After initializing the chain by drawing from an arbitrary distribution $x_0 \sim \pi_0(x)$, the updating rule writes

$$x_{k+1} \leftarrow x_k + \frac{\epsilon^2}{2} \nabla_x \log q_X(x_k) + \epsilon \xi \quad (7)$$

where $\xi \sim \mathcal{N}(0, I)$. When $\epsilon \rightarrow 0$ and the number of samples $K \rightarrow \infty$, the distribution of the samples generated by the iterative procedure (7) converges to q_X under some regularity conditions. In practice, the error is negligible when ϵ is sufficiently small and K is sufficiently large. This algorithm referred to as the unadjusted Langevin Algorithm (ULA) always accepts the generated sample proposed by (7), neglecting the errors induced by the discretization scheme of the continuous diffusion. To correct this bias, Metropolis-adjusted Langevin Algorithm (MALA) applies a Metropolis-Hastings step to accept or reject a sample proposed by ULA [Grenander and Miller, 1994].

Again, sampling according to q_X thanks to the diffusion (7) is likely to be inefficient due to the expected complexity of the target distribution possibly defined over a subspace of \mathbb{R}^d . In particular, this strategy suffers from the lack of prior knowledge about the location of the mass. Conversely, the proposed approach explores the latent space by leveraging on the closed-form change of variable (2) operated by the trained NF. After technical derivations reported in Appendix

C.2, the counterpart of the diffusion (7) expressed in the latent space writes

$$z' = z_k + \frac{\epsilon^2}{2} G^{-1}(z_k) \nabla_z \log \tilde{q}_Z(z_k) + \epsilon \sqrt{G^{-1}(z_k)} \xi \quad (8)$$

where

$$\tilde{q}_Z(z) = q_Z(z) |J_f(z)|^{-1} \quad (9)$$

and

$$G^{-1}(z) = [J_f^{-1}(z)]^2. \quad (10)$$

Note that the distribution \tilde{q}_Z in (9) originates from the change of variable that defines q_X in (2) and has been already implicitly introduced by (6) in Section 4. Interestingly, the matrix $G(\cdot)$ is a definite positive matrix (see Appendix B). Thus the diffusion (8) characterizes a Riemannian manifold Langevin dynamics where $G(\cdot)$ is the Riemannian metric associated with the latent space [Girolami and Calderhead, 2011, Xifara et al., 2014]. More precisely, it defines the conventional proposal move of the Riemannian manifold adjusted Langevin algorithm (RMMALA) which targets the distribution \tilde{q}_Z defined by (9). This distribution is explicitly defined through the Jacobian $J_f(\cdot)$ of the transformation whose behavior has been discussed in depth in Section 4. It can be interpreted as the Gaussian latent distribution q_Z tempered by the (determinant of the) Jacobian of the transformation. It has also been evidenced by depicting the heatmap of (6) in Fig. 1(b), which shows that it appears as a better approximation of p_Z than q_Z . Since it governs the drift of the diffusion through the gradient of its logarithm, the diffusion is expected to escape from the areas where the determinant of the Jacobian explodes, see Section 4.

The proposal kernel $g(z'|z)$ associated with the diffusion (8) is a Gaussian distribution whose probability density function (pdf) can be conveniently rewritten as (see Property 5 in Appendix C.2)

$$g(z' | z_k) \propto |J_f(z_k)| \exp \left[-\frac{1}{2\epsilon^2} \left\| J_f(z_k)(z' - z_k) + \frac{\epsilon^2}{2} \tilde{s}_Z(z_k) \right\|^2 \right]. \quad (11)$$

where $\tilde{s}_Z(\cdot)$ denotes the so-called latent score

$$\tilde{s}_Z(z) = J^{-1}(z) \nabla_z \log \tilde{q}_Z(z). \quad (12)$$

The sample proposed according to (8) is then accepted with probability

$$\alpha_{\text{RMMALA}}(z_k, z') = \min \left(1, \frac{\tilde{q}_Z(z') g(z_k | z')}{\tilde{q}_Z(z_k) g(z' | z_k)} \right). \quad (13)$$

It is worth noting that the formulation (11) of the proposal kernel leads to a significantly faster implementation than its canonical formulation. Indeed, it does not require to compute the metric $G^{-1}(\cdot)$ defined by (10), which depends on the inverse of the Jacobian matrix twice. Moreover, the evaluation of the latent score (12) can be achieved in an efficient manner, bypassing the need for evaluating the inverse of the Jacobian matrix, as elaborated in Appendix C.3.2. Finally, only the Jacobian associated with the forward transformation is required to compute (11). This approach enables a streamlined calculation of the acceptance ratio (13), ensuring an overall computational efficiency.

Besides, the proposal scheme (8) requires to generate high dimensional Gaussian variables with covariance matrix $\epsilon^2 G^{-1}(\cdot)$ [Vono et al., 2022]. To lighten the corresponding computational burden, we take advantage of a 1st order expansion of f^{-1} to approximate (8) by the diffusion (see Appendix C.3.1)

$$z' = f^{-1}(f(z_k) + \epsilon \xi) + \frac{\epsilon^2}{2} J_f^{-1}(z_k) \tilde{s}_Z(z). \quad (14)$$

According to (14), this alternative proposal scheme requires to generate high dimensional Gaussian variables with a covariance matrix which is now identity, i.e., most cheaper. Moreover, it is worth noting that *i*) the latent score $\tilde{s}_Z(\cdot)$ can be evaluated efficiently (see above) and *ii*) using $J_f^{-1}(z) = J_{f^{-1}}(f(z))$ (see Property 2 in Appendix C.1), sampling z' according to (14) only requires to evaluate the Jacobian associated with the backward transformation. Proofs and implementation details are reported in Appendix C. The algorithmic procedure to sample according to this kernel denoted $\mathcal{K}_{\text{RMMALA}}(\cdot)$ is summarized in Algo. 1.

5.2 Independent Metropolis-Hastings sampling

Handling distributions that exhibit several modes or defined on a complex multi-component topology is another major issue raised by the problem addressed here. In practice, conventional sampling schemes such as those based on Langevin dynamics fail to explore the full distribution when modes are isolated since they may get stuck around one

Algorithm 1: Sampling kernel $\mathcal{K}_{\text{RMMALA}}(\cdot)$.

Input: trained NF $f(\cdot)$, time step ϵ , current state z_k of the sampler.
/* Draw the candidate */
1 Draw $\xi \sim \mathcal{N}(0, 1)$
2 Set $z' = f^{-1}(f(z_k) + \epsilon \cdot \xi) + \frac{\epsilon^2}{2} J_f^{-1}(z_k) \tilde{s}_Z(z_k)$
/* Accept/reject procedure */
3 Draw $u \sim \mathcal{U}(0, 1)$
4 **if** $u < \alpha_{\text{RMMALA}}(z_k, z')$ **then**
5 | Set $z_{k+1} = z'$
6 **else**
7 | Set $z_{k+1} = z_k$
Output: New state $z_{k+1} = \mathcal{K}_{\text{RMMALA}}(z_k)$ of the sampler.

of these modes. Thus, the samples proposed according to (8) in areas with high values of $\|J_f(\cdot)\|$ are expected to be rejected. These areas have been identified in Section 4 as the low probability regions between modes when targeting a multimodal distribution. To alleviate this problem, one strategy consists in resorting to another kernel to propose moves from one high probability region to another, without requiring to cross the low probability regions. Following this strategy, this paper proposes to combine the diffusion (8) with an independent Metropolis-Hastings (I-MH) with the distribution q_Z as a proposal. The corresponding acceptance ratio writes

$$\begin{aligned}\alpha_{\text{I-MH}}(z_k, z') &= \min \left(1, \frac{\tilde{q}_Z(z') q_Z(z_k)}{\tilde{q}_Z(z_k) q_Z(z')} \right) \\ &= \min \left(1, \frac{|J_f(z_k)|}{|J_f(z')|} \right).\end{aligned}\tag{15}$$

It is worth noting that this probability of accepting the proposed move only depends on the ratio between the Jacobians evaluated at the current and the candidate states. In particular, candidates located in regions of the latent space characterized by exploding Jacobians in case of a topological mismatch (see Section 4) are expected to be rejected with high probability. Conversely, this kernel will favor moves towards other high probability regions not necessarily connected to the regions of the current state. The algorithmic procedure is sketched in Algo. 2.

Algorithm 2: Sampling kernel $\mathcal{K}_{\text{I-MH}}(\cdot)$.

Input: trained NF $f(\cdot)$, current state z_k of the sampler.
/* Draw candidate */
1 Draw $z' \sim \mathcal{N}(0, 1)$
/* Accept/reject procedure */
2 Draw $u \sim \mathcal{U}(0, 1)$
3 **if** $u < \alpha_{\text{I-MH}}(z_k, z')$ **then**
4 | Set $z_{k+1} = z'$
5 **else**
6 | Set $z_{k+1} = z_k$
Output: New state $z_{k+1} = \mathcal{K}_{\text{I-MH}}(z_k)$ of the sampler.

Finally, the overall proposed sampler, referred to as NF-SAILS for NF SAmpling In the Latent Space and summarized in Algo. 3, combines the transition kernels $\mathcal{K}_{\text{RMMALA}}$ and $\mathcal{K}_{\text{I-MH}}$, which permits to efficiently explore the latent space both locally and globally. At each iteration k of the sampler, the RMMALA kernel $\mathcal{K}_{\text{RMMALA}}$ associated with the acceptance ratio (13) is selected with probability p and the I-MH kernel $\mathcal{K}_{\text{I-MH}}$ associated with acceptance ratio (15) is selected with the probability $1 - p$. Again, one would like to emphasize that the proposed strategy does not depend on the NF architecture and can be adopted to sample from any pretrained NF model.

6 Experiments

This section reports performance results to illustrate the efficiency of NF-SAILS thanks to experiments based on several models and synthetic data sets. It is compared to state-of-the-art generative models known for their abilities to

Algorithm 3: NF-SAILS: NF SAmpling In the Latent Space.

Input: trained NF $f(\cdot)$, time step ϵ , probability p
/* Initialization */
1 Draw $z_0 \sim \pi_0(z)$
2 **for** $k = 0$ **to** K **do**
 /* Choose the kernel */
 3 Draw $u \sim \mathcal{U}(0, 1)$
 4 **if** $u < p$ **then**
 /* LOCAL EXPLORATION (see Algo. 1) */
 5 $z_{k+1} = \mathcal{K}_{\text{RMMALA}}(z_k)$
 6 **else**
 /* GLOBAL EXPLORATION (see Algo. 2) */
 7 $z_{k+1} = \mathcal{K}_{\text{I-MH}}(z_k)$
8 **end**

Output: Collection of samples $\{z_k\}_{k=1}^K$.

handle multimodal distributions. These results will show that the proposed sampling strategy achieves good performance, without requiring to adapt the NF training procedure or resorting to non-Gaussian latent distributions. We will also confirm the relevance of the method when working on popular image data sets, namely Cifar-10 [Krizhevsky et al., 2010], CelebA [Liu et al., 2015] and LSUN Yu et al. [2015].

To illustrate the versatility of proposed approach w.r.t. the NF architecture, two types of coupling layers are used to build the trained NFs. For the experiments conducted on the synthetic data sets, the NF architecture is RealNVP [Dinh et al., 2016]. Conversely, a Glow model is used for experiments conducted on the image data sets [Kingma and Dhariwal, 2018]. However, it is worth noting that the proposed method can apply on top of any generative model fitting multimodal distributions. Additional details regarding the training procedure are reported in Appendix D.1.

6.1 Figures-of-merit

To evaluate the performance of the NFs, several figures-of-merit have been considered. When addressing bi-dimensional problems, we perform a Kolmogorov-Smirnov test to assess the quality of the generated samples w.r.t. the underlying true target distribution [Justel et al., 1997]. The goodness-of-fit is also monitored by evaluating the mean log-likelihood of the generated samples and the entropy estimator between samples, which approximates the Kullback-Leibler divergence between empirical samples [Kraskov et al., 2004].

For applications to higher dimensional problems, such as image generation, the performances of the compared algorithms are evaluated using the Fréchet inception distance (FID) [Heusel et al., 2017] using a classifier pre-trained specifically on each data set. Besides, for completeness, we report the bits per dimension (bpd) [Papamakarios et al., 2017], i.e., the log-likelihoods in the logit space, since this is the objective optimized by the trained models.

6.2 Results obtained on synthetic data set

As a first illustration of the performance of NF-SAILS, we consider to learn a mixture of k bidimensional Gaussian distributions, with $k \in \{2, 3, 4, 6, 9\}$. The NF model $f(\cdot)$ is a RealNVP [Dinh et al., 2016] composed of $M = 4$ flows, each composed of two three-layer neural networks ($d \rightarrow 16 \rightarrow 16 \rightarrow d$) using hyperbolic tangent activation function. We use the Adam optimizer with learning rate 10^{-4} and a batch size of 500 samples.

Table 1 reports the considered metrics when comparing the proposed NF-SAILS sampling method to a naive sampling (see Section 3.3) or to state-of-the-art sampling techniques from the literature, namely Wasserstein GAN with gradient penalty (WGAN-GP) [Gulrajani et al., 2017] and denoising diffusion probabilistic models (DDPM) [Ho et al., 2020]. These results show that NF-SAILS consistently competes favorably against the compared methods, in particular as the degree of multimodality of the distribution increases. Note that WGAN-GP exploits a GAN architecture. Thus, contrary to the proposed NF-based sampling method, it is unable to provide an explicit evaluation of the likelihood, which explains the N/A values in the table.

Figure 3 illustrates this result for $k = 6$ and shows that our method considerably reduces the number of out-of-distribution generated samples. Additional results are reported in Appendix D.2.

		$\uparrow \log p_X$	$\downarrow \text{KL}$	$\downarrow \text{KS}$
$k = 2$	Naive sampling	-4.13	0.263	0.177
	NF-SAILS	-1.41	0.057	0.047
	WGAN-GP	N/A	0.308	0.287
	DDPM	-2.98	0.121	0.066
$k = 3$	Naive sampling	-3.52	0.914	0.248
	NF-SAILS	-1.83	0.056	0.034
	WGAN-GP	N/A	0.973	0.237
	DDPM	-2.97	0.364	0.124
$k = 4$	Naive sampling	-3.08	0.967	0.295
	NF-SAILS	-1.07	0.044	0.041
	WGAN-GP	N/A	1.012	0.317
	DDPM	-1.81	0.427	0.127
$k = 6$	Naive sampling	-2.06	1.219	0.205
	NF-SAILS	-1.09	0.039	0.309
	WGAN-GP	N/A	1.392	0.212
	DDPM	-1.99	1.004	0.179
$k = 9$	Naive sampling	-2.297	1.764	0.215
	NF-SAILS	-0.801	0.151	0.052
	WGAN-GP	N/A	1.939	0.340
	DDPM	-1.258	0.906	0.205

Table 1: Goodness-of-fit of the generated samples w.r.t. the number k of Gaussians. Reported scores result from the average over 5 Monte Carlo runs.

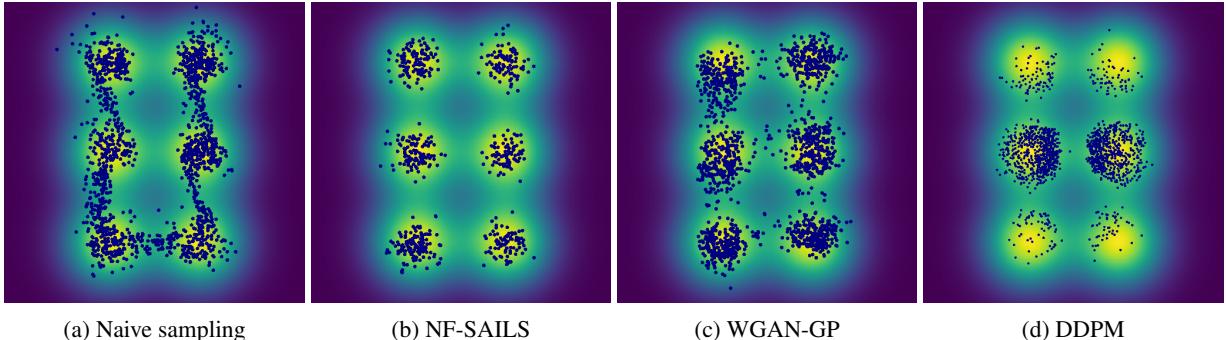


Figure 3: Mixture of $k = 6$ Gaussian distributions (green), and 1000 generated samples (blue). The proposed NF-SAILS method in Fig. 3b does not generate samples in-between modes.

6.3 Results obtained on real image data sets

Moreover, we further study the performance of NF-SAILS on three different real image data sets, namely Cifar-10 [Krizhevsky et al., 2010], CelebA [Liu et al., 2015] and LSUN [Yu et al., 2015]. Following the same protocol as implemented by Kingma and Dhariwal [2018], we use a Glow architecture where each neural network are composed of three convolutional layers. The two hidden layers have ReLU activation functions and 512 channels. The first and last convolutions are 3×3 , while the center convolution is 1×1 , since its input and output have a large number of channels, in contrast with the first and last convolutions. Details regarding the implementation are reported in Appendix D.3.

We compare the FID score as well as the average negative log-likelihood (bpd), keeping all training conditions constant and averaging the results over 10 Monte Carlo runs. The results are depicted in Fig. 4 reports the results when compared to those obtained by naive sampling or WGAN-GP [Gulrajani et al., 2017]. As shown by the different panels of this figure, the proposed NF-SAILS method considerably improves the quality of the generated images, both quantitatively (in term of FID) and semantically. Our methodology compares favourably w.r.t. to WGAN-GP for the two data sets CelebA and LSUN.

	\downarrow bpd	\downarrow FID
Naive sampling	3.35	44.6
NF-SAILS	3.08	43.11
WGAN-GP	N/A	18.8



(a) Cifar10

	\downarrow bpd	\downarrow FID
Naive sampling	1.03	15.82
NF-SAILS	0.96	14.12
WGAN-GP	N/A	12.89



(b) CelebA

	\downarrow bpd	\downarrow FID
Naive sampling	2.38	8.91
NF-SAILS	2.11	7.86
WGAN-GP	N/A	9.56



(c) LSUN (bedroom)

Figure 4: Tables report quantitative and perceptual metrics computed from the samples generated by the compared methods. The figures show some samples generated from Glow using the proposed NF-SAILS method.

7 Conclusion

This paper discusses the sampling from the target distribution learnt by a normalizing flow. Architectural constraints prevent normalizing flows to properly learn disconnect support measures due to the topological mismatch between the latent and target spaces. Moreover, we theoretically prove that Jacobian norm of the transformation become arbitrarily large to closely represent such target measures. The conducted analysis exhibits the existence of pathological areas in the latent space corresponding to points with exploding Jacobian norms. Using a naive sampling strategy leads to out of distribution samples located in these areas. To overcome this issue, we propose a new sampling procedure based on a Langevin diffusion directly formulated in the latent space. This sampling is interpreted as a Riemannian manifold Metropolis adjusted Langevin algorithm, whose metric is driven by the Jacobian of the learnt transformation. This local exploration of the latent space is complemented by an independent Metropolis-Hastings kernel which allows moves from one high probability region to another while avoiding crossing pathological areas. One particular advantage of the proposed is that it can be applied to any pre-trained NF model. Indeed it does not require a particular training strategy of the NF or to adapt the distribution assumed in the latent space. The performances of the proposed sampling strategy show to compare favorably to state-of-the art, with very few out-of-distribution samples.

Appendix A Proof of Proposition 1

The proof of Proposition 1 in Section 4.1 combines existing results from topology and real analysis. The complete background can be found in [Dudley, 2002] and [Cornish et al., 2020]. The proof is mainly based on the following results.

Theorem 1 (Cornish et al. [2020]). *Let q_Z and q_X define probability measures on \mathbb{R}^d , with $\text{supp}(q_Z) \neq \text{supp}(q_X)$. For any sequence of measurable, differentiable Lipschitzian functions $f_n : \mathbb{R}^d \rightarrow \mathbb{R}^d$, if the sequence weakly converges as $f_n \# q_Z \xrightarrow{\mathcal{D}} q_X$, then*

$$\lim_{n \rightarrow \infty} \text{Lip } f_n = \infty. \quad (16)$$

Moreover, Behrmann et al. [2019] showed the relation between the Lipschitz constant and the Jacobian of a transformation, as stated below.

Lemma 1 (Rademacher’s theorem). *If $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is Lipschitzian, then f is continuous and differentiable at almost all points of \mathbb{R}^m and*

$$\text{Lip } f = \sup_{z \in \mathcal{Z}} \|J_f(z)\|_{\text{op}} \quad (17)$$

Both Theorem 1 and Lemma 1 rely on the same starting hypothesis, i.e., f is required to be continuous, differentiable and Lipschitzian. Combining these two results yields Proposition 1 following a development of the proof of the results by Cornish et al. [2020] and Behrmann et al. [2019].

Appendix B Properties of the Jacobian of coupling layer-based NFs

B.1 Structure of the Jacobian matrix and computation of its determinant

RealNVP model defines a NF by implementing a sequence of M invertible bijective transformation functions, herein referred to as coupling layers [Dudley, 2002]. In other words, the mapping f writes as $f = f^{(M)} \circ f^{(M-1)} \circ f^{(2)} \circ f^{(1)}$. Each bijection $f^{(m)} : u \mapsto v$ associated to the m th layer splits the input $u \in \mathbb{R}^D$ into two parts of sizes d and $d - D$ ($d \leq D$), respectively, such that the output $v \in \mathbb{R}^D$ writes

$$\begin{cases} v_{1:d} &= u_{1:d} \\ v_{d+1:D} &= u_{d+1:D} \odot \exp(h^{(m)}(u_{1:d})) + t^{(m)}(u_{1:d}) \end{cases} \quad (18)$$

where $h_m(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{D-d}$ and $t_m(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{D-d}$ are scale and translation functions implemented as deep networks and \odot stands for the Hadamard product. The Jacobian of the above transformation is a lower triangular matrix

$$J^{(m)}(u) = \begin{bmatrix} I_d & \mathbf{0}_{d \times (D-d)} \\ A^{(m)}(u) & E^{(m)}(u) \end{bmatrix} \quad (19)$$

where I_d and $\mathbf{0}_{d \times (D-d)}$ are the identity and zero matrices with indexed sizes, respectively, and

$$\begin{cases} A^{(m)}(u) &= u_{d+1:D} \odot \frac{\partial \exp h^{(m)}(u_{1:d})}{\partial u_{1:d}} + \frac{\partial t^{(m)}(u_{1:d})}{\partial u_{1:d}} \\ E^{(m)}(u) &= \text{diag}(\exp(h^{(m)}(u_{1:d}))) . \end{cases} \quad (20)$$

Thanks to the chain rule, it follows that the Jacobian of the overall NF is

$$J_f(z) = \prod_{j=1}^J J^{(m)}(u^{(m)}) \quad (21)$$

with $u^{(m)} = f^{(m-1)}(u^{(m-1)})$ and $z = u^{(0)}$.

Moreover, because of the structure of each layer, the determinant of the Jacobian $J^{(m)}(u)$ associated with the m th layer is

$$|J^{(m)}(u)| = \prod_{k=1}^d \exp(h^{(m)}(u_k)) . \quad (22)$$

The determinant of the Jacobian $J_f(\cdot)$ characterizing the overall NF can be easily computed from (21) and (22).

B.2 Positive definiteness of the Jacobian

Property 1. *The product of two lower triangular matrices with strictly positive diagonal elements is a positive definite lower triangular matrix.*

Proof. Let $A = [a_{ij}]$ and $B = [b_{ij}]$ be two $n \times n$ lower triangular matrices with positive diagonal entries, i.e.,

$$\forall i, j \quad \text{such that } i < j, \text{ then } a_{ij} = b_{ij} = 0. \quad (23)$$

$$\forall i \quad a_{ii} > 0 \text{ and } b_{ii} > 0 \quad (24)$$

Let $C = [c_{ij}]$ denote the product matrix $C = AB$ with $c_{ij} = \sum_{k=1}^n a_{ik}b_{kj}$. The upper elements c_{ij} ($i < j$) of C can be computed as

$$c_{ij} = \sum_{k=1}^i a_{ik}b_{kj} + \sum_{k=i+1}^n a_{ik}b_{kj}. \quad (25)$$

In the right hand side of (25), if $k \leq i$ then $b_{kj}=0$. Moreover if $k > i$ then $a_{ik} = 0$. As a consequence, $c_{ij} = 0$ and C is triangular.

Moreover, the eigenvalues of a triangular matrix is its diagonal elements. It follows that C is positive definite. \blacksquare

Thanks to the structure of coupling layer-based NFs discussed in Appendix B.1, we have the two following corollaries.

Corollary 1. *The Jacobian matrix $J_f(\cdot)$ and its inverse $J_f^{-1}(\cdot)$ of coupling layer-based NFs are positive definite.*

Corollary 2. *The matrix $G(\cdot)$ and its inverse $G^{-1}(\cdot)$ are positive definite.*

Appendix C Diffusion in the latent space

C.1 Preliminaries

The Langevin diffusion is a particular instance of the Itô process defined in the following Lemma of which a proof is given in [Øksendal, 2003].

Lemma 2 (Itô's lemma). *Let X_t denote an Itô drift-diffusion process defined by the stochastic differential equation*

$$dX_t = \mu_t dt + \sigma_t dB_t. \quad (26)$$

If $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a differentiable scalar function, then

$$df(t, X_t) = \left(\frac{\partial f}{\partial t} + \mu_t \frac{\partial f}{\partial x} + \frac{\sigma_t^2}{2} \frac{\partial^2 f}{\partial x^2} \right) dt + \sigma_t \frac{\partial f}{\partial x} dB_t. \quad (27)$$

It yields that $f(t, X_t)$ is an Itô drift-diffusion process itself.

The following property shows that for any bijective transformation, the Jacobian of the inverse transformation is equal to the inverse of the Jacobian of the transformation. This result will be useful later.

Property 2. *Let $f : \mathcal{Z} \rightarrow \mathcal{X}$ denote a bijective transformation and $J_f(\cdot)$ its Jacobian, then*

$$J_{f^{-1}}(f(z)) = J_f^{-1}(z).$$

Proof. Let h and g denote two multivariate functions. The chain rule writes

$$J_{h \circ g}(\cdot) = J_h(g(\cdot)) J_g(\cdot) \quad (28)$$

thus

$$J_h(g(\cdot)) = J_{h \circ g}(\cdot) J_g^{-1}(\cdot). \quad (29)$$

Moreover, for any multivariate bijective function f , we have

$$J_{f \circ f^{-1}}(\cdot) = J_{f^{-1} \circ f}(\cdot) = I_d. \quad (30)$$

Combining (29) and (30) with $h = f^{-1}$ and $g = f$ yields

$$J_{f^{-1}}(f(z)) = J_{f^{-1} \circ f}(z) J_f^{-1}(z) = I_d J_f^{-1}(z) = J_f^{-1}(z). \quad (31)$$

■

The following property demonstrates that the gradient of the score of q_X can be expressed over the latent space \mathcal{Z} using \tilde{q}_Z defined in (9).

Property 3. *Let $f : \mathcal{Z} \rightarrow \mathcal{X}$ be a bijective transformation which maps a latent measure q_Z towards a target measure q_X . Then the score of $q_X(x)$ is given by*

$$\nabla_x \log q_X(x) = J_f^{-1}(z) \cdot \nabla_z \log \tilde{q}_Z(z)$$

where $\tilde{q}_Z(z) = q_Z(z) |J_f(z)|^{-1}$.

Proof. From the definition of $q_X(x)$ in equation (2), the score of $q_X(x)$ writes

$$\nabla_x \log q_X(x) = \nabla_x [\log q_Z(f^{-1}(x)) + \log |J_{f^{-1}}(x)|] \quad (32)$$

and, from Property 2,

$$\nabla_x \log q_X(x) = \nabla_{f(z)} [\log q_Z(z) + \log |J_f(z)|^{-1}] \quad (33)$$

$$= \nabla_{f(z)} \log \tilde{q}_Z(z) \quad (34)$$

The chain rule now leads to

$$\nabla_x \log q_X(x) = \nabla_x f^{-1}(x) \cdot \nabla_z \log \tilde{q}_Z(z) \quad (35)$$

$$= J_{f^{-1}}(f(z)) \cdot \nabla_z \log \tilde{q}_Z(z) \quad (36)$$

which, using Property 2, can be finally rewritten as

$$\nabla_x \log q_X(x) = J_f^{-1}(z) \cdot \nabla_z \log \tilde{q}_Z(z). \quad (37)$$

■

C.2 Derivation of the proposal distribution

The following property shows that the Lanvegin diffusion which targets the distribution q_X can be rewritten as a diffusion over the latent space \mathcal{Z} .

Property 4. *We consider the overdamped Langevin Itô diffusion*

$$dX_t = \nabla_x \log q_X(X_t) dt + \sigma_t dB_t \quad (38)$$

driven by the time derivative of a standard Brownian motion B_t . In the limit $t \rightarrow \infty$, this probability distribution X_t approaches a stationary distribution q_X . Let $f : \mathcal{Z} \rightarrow \mathcal{X}$ be a bijective transformation which maps a latent measure q_Z towards the target measure q_X . A counterpart Langevin diffusion expressed over the latent space \mathcal{Z} writes

$$dZ_t = G^{-1}(Z_t) \nabla_z \log \tilde{q}_Z(Z_t) dt + \sigma_t \sqrt{G^{-1}(Z_t)} dB_t \quad (39)$$

Proof. The Langevin diffusion is a particular instance of the Itô process where the drift μ_t in (26) is given by the gradient of the log-density $\nabla_x \log q_X(X_t)$, i.e.,

$$dX_t = \nabla_x \log q_X(X_t) dt + \sigma_t dB_t \quad (40)$$

We are interested in the diffusion process of $f^{-1}(X_t)$ when $f(\cdot)$ is a NF which is continuous, differentiable and bijective such that $f(Z_t) = X_t$ and $f^{-1}(X_t) = Z_t$. The Ito's Lemma 2 states

$$df^{-1}(X_t) = \left(J_{f^{-1}}(X_t) \nabla_x \log q_X(X_t) + \frac{\sigma_t^2}{2} \text{tr}(H_{f^{-1}}(X_t)) \right) dt + \sigma_t J_{f^{-1}}(X_t) dB_t. \quad (41)$$

Neglecting the second-order terms yields

$$df^{-1}(X_t) = J_{f^{-1}}(X_t) \nabla_x \log q_X(X_t) dt + \sigma_t J_{f^{-1}}(X_t) dB_t. \quad (42)$$

Using Property 2, Eq. (42) can be rewritten as

$$dZ_t = J_f^{-1}(Z_t) \nabla_x \log q_X(X_t) dt + \sigma_t J_f^{-1}(Z_t) dB_t. \quad (43)$$

Finally, by denoting $G^{-1}(z) = [J_f^{-1}(z)]^2$ and using Property 3, the diffusion in the latent space writes

$$dZ_t = G^{-1}(Z_t) \nabla_z \log \tilde{q}_Z(Z_t) dt + \sigma_t \sqrt{G^{-1}(Z_t)} dB_t \quad (44)$$

■

The discretization of the stochastic differential equation (44) using the Euler–Maruyama scheme can be written as in (8). This discretized counterpart of the diffusion corresponds to the proposal move of a Riemann manifold Metropolis–Adjusted Langevin algorithm which targets \tilde{q}_Z . The following property shows that the associated proposal kernel can be rewritten as (11).

Property 5. *The discrete Langevin diffusion given by*

$$z' = z + \frac{\epsilon^2}{2} \cdot G^{-1}(z) \nabla_z \log \tilde{q}_Z(z) + \epsilon \cdot \sqrt{G^{-1}(z)} \xi \quad (45)$$

with $\xi \sim \mathcal{N}(0, I)$ is defined by the transition kernel

$$q(z' | z) \propto |J_f(z)| \exp \left[-\frac{1}{2\epsilon^2} \left\| J_f(z)(z' - z) + \frac{\epsilon^2}{2} J_f^{-1}(z) \nabla_z \log \tilde{q}_Z(z) \right\|^2 \right]. \quad (46)$$

Proof. From the Gaussian nature of ξ , the conditional distribution of z' is a Gaussian distribution whose mean is governed by the drift and covariance matrix is parametrized by the (inverse of the Jacobian), namely

$$z' | z \sim \mathcal{N}(\mu, \Sigma) \quad (47)$$

with

$$\mu = z + \frac{\epsilon^2}{2} \cdot G^{-1}(z) \nabla_z \log \tilde{q}_Z(z) \quad (48)$$

$$\Sigma = \epsilon^2 J_f^{-1}(z) J_f^{-\top}. \quad (49)$$

The corresponding pdf writes

$$q(z' | z) = \left(\frac{1}{2\pi} \right)^{\frac{d}{2}} \frac{1}{|\Sigma|^{1/2}} \exp \left(-\frac{1}{2}(z' - \mu)^\top \Sigma^{-1} (z' - \mu) \right). \quad (50)$$

First, let notice that $\Sigma^{-1} = \epsilon^{-2} J_f^\top(z) J_f(z)$. Then we have

$$(z' - \mu)^\top \Sigma^{-1} (z' - \mu) = \epsilon^{-2} \left[z' - z - \frac{\epsilon^2}{2} \cdot [J_f^{-1}(z)]^2 \nabla_z \log \tilde{q}_Z(z) \right]^\top J_f^\top(z) \quad (51)$$

$$\times J_f(z) \left[z' - z - \frac{\epsilon^2}{2} \cdot [J_f^{-1}(z)]^2 \nabla_z \log \tilde{q}_Z(z) \right]$$

$$= \epsilon^{-2} \left\| J_f(z)(z' - z) - \frac{\epsilon^2}{2} J_f^{-1}(z) \nabla_z \log \tilde{q}(z) \right\|^2 \quad (52)$$

Finally, using $|\Sigma|^{1/2} = \epsilon |J_f^{-1}(z)| = \epsilon |J_f(z)|^{-1}$ yields

$$q(z' | z) \propto |J_f(z)| \exp \left[-\frac{1}{2\epsilon^2} \left\| J_f(z)(z' - z) - \frac{\epsilon^2}{2} J_f^{-1}(z) \nabla_z \log \tilde{q}(z) \right\|^2 \right]. \quad (53)$$

■

It is worth noting that the pdf of this transition kernel should be computed when evaluating the acceptance ratio (13). When using the canonical writing (50), evaluating this pdf would require to compute $G^{-1}(z)$ in (48) and $J_f^{-1}(z) J_f^{-\top}(z)$ in (49). Instead, evaluating this pdf with the specific form (53) only requires to compute $J_f(z)$ since the latent score $\tilde{s}_Z(z) = J_f^{-1}(z) \nabla_z \log \tilde{q}(z)$ can be computed efficiently, as discussed later in Appendix C.3.2.

C.3 Efficient implementation

C.3.1 Approximation of the proposal move

Generating high dimension Gaussian variables according to (8) is expected to be very costly because of the covariance matrix, even if the corresponding Cholesky factor $\epsilon J_f^{-1}(\cdot)$ is lower triangular and explicit (see Appendix B.1). Alternatively, to lighten the computation, we take advantage of the 1st order expansion

$$f^{-1}(f(z_k) + \epsilon \xi) \simeq f^{-1} \circ f(z_k) + \epsilon J_{f^{-1}}(f(z_k)) \xi. \quad (54)$$

Using Property 2, this amounts to approximate (8) by the diffusion

$$z' = f^{-1}(f(z_k) + \epsilon \xi) + \frac{\epsilon^2}{2} G^{-1}(z_k) \nabla_z \log \tilde{q}(z) \quad (55)$$

$$= f^{-1}(f(z_k) + \epsilon \xi) + \frac{\epsilon^2}{2} J_f^{-1}(z_k) \tilde{s}_Z(z). \quad (56)$$

According to (56), this alternative proposal scheme only requires to generate high dimensional Gaussian variables whose covariance matrix is now identity.

C.3.2 Fast computation of the latent score

The latent score $\tilde{s}_Z(z)$ is a critical quantity in the proposed method, as it contributes to the drift term in the proposal move (14) and to the proposal kernel (11). Property 3 shows that the latent score is equal to the score of q_X expressed in the target domain, i.e.,

$$\nabla_x \log q_X(x) = J_f^{-1}(z) \cdot \nabla_z \log \tilde{q}_Z(z)$$

The adopted implementation bypasses the costly evaluation and storage of the inverse Jacobian by directly computing the latent score as $\nabla_x \log q_X(x)$. The evaluation of the score of q_X can be conveniently performed thanks to the auto-differentiation modules provided by numerous deep learning frameworks.

Appendix D Experiments

D.1 Training

In all experiments we trained our models to maximize either the log-likelihood using the ADAM optimiser with default hyperparameters and no weight decay. We used a held-out validation set and trained each model until its validation score stopped improving, except for the synthetic data experiments where we train for a fixed number of 1000 epochs.

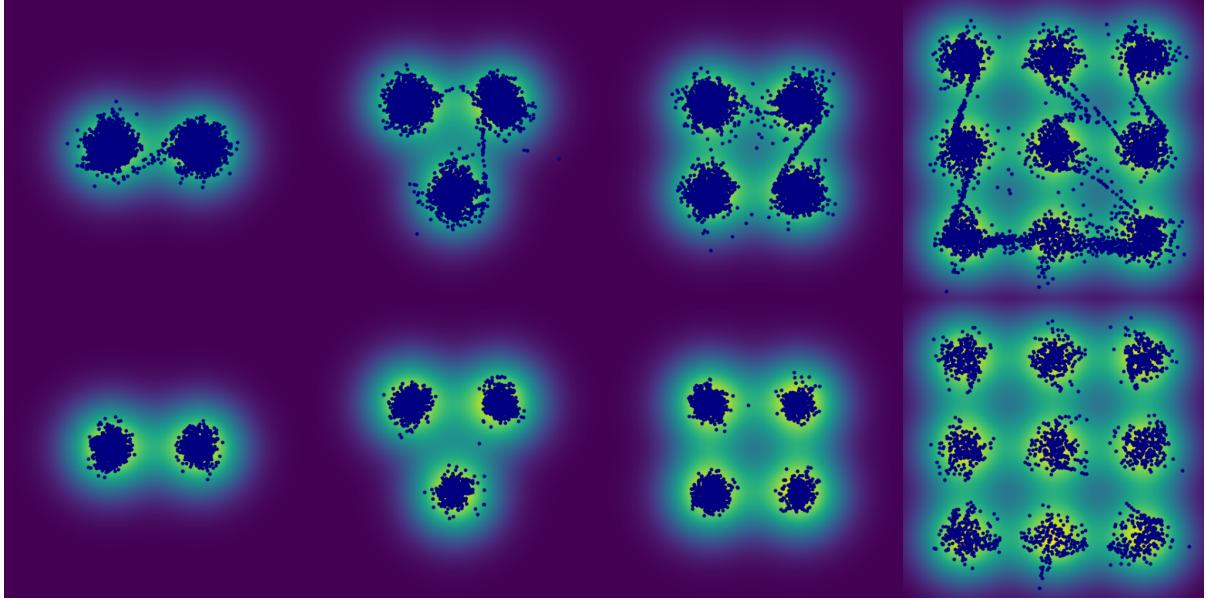


Figure 5: Mixture of k Gaussian distributions (green), and 1000 samples (blue) generated by the naive sampling (top) and the proposed NF-SAILS method (bottom) with, from left to right, $k = 2, k = 3, k = 4$ and $k = 9$.

D.2 Complementary results for the synthetic experiments

Figure 5 shows the difference of sampling quality between naive sampling and the proposed NF-SAILS method for RealNVP model trained on k -mixtures of Gaussians for $k \in \{2, 3, 4, 9\}$. See also Table 1 in Section 6.2 of the main document.

D.3 Implementation details for the image experiments

The hyperparameters used in the experiments conducted on images (see Section 6.3 of the main document) are reported in Table 2.

Dataset	Minibatch Size	Levels (L)	Depth per level (K)	Coupling
CIFAR-10	512	3	32	Affine
LSUN, 64×64	128	4	48	Affine
LSUN, 96×96	320	5	64	Affine
LSUN, 128×128	160	5	64	Affine
CelebA, 96×96	320	5	64	Affine
CelebA, 128×128	160	6	32	Affine

Table 2: Architectures of the Glow model implemented for the experiments conducted on the image data sets.

References

- L. Ardizzone, R. Mackowiak, C. Rother, and U. Köthe. Training normalizing flows with the information bottleneck for competitive generative classification. In *Adv. in Neural Information Process. Systems (NeurIPS)*, 2020.
- M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proc. Int. Conf. Machine Learning (ICML)*, Proceedings of Machine Learning Research. PMLR, 2017.
- G. Arvanitidis, L. K. Hansen, and S. Hauberg. Latent space oddity: on the curvature of deep generative models. In *Proc. IEEE Int. Conf. Learn. Represent. (ICLR)*, 2018.
- J. Behrmann, P. Vicol, K.-C. Wang, R. B. Grosse, and J.-H. Jacobsen. On the invertibility of invertible neural networks, 2019. URL <https://openreview.net/forum?id=BJ1VeyHFwH>.

- Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Patt. Anal. Mach. Intell.*, 2013.
- F. Coeurdoux, N. Dobigeon, and P. Chainais. Sliced-Wasserstein normalizing flows: beyond maximum likelihood training. In *Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Belgium, Oct. 2022.
- R. Cornish, A. Caterini, G. Deligiannidis, and A. Doucet. Relaxing bijectivity constraints with continuously indexed normalising flows. In *Proc. Int. Conf. Machine Learning (ICML)*, pages 2133–2143. PMLR, 2020.
- L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real NVP. *arXiv preprint arXiv:1605.08803*, 2016.
- R. M. Dudley. *Real analysis and probability*. Cambridge University Press, 2002.
- M. Gabrié, G. M. Rotskoff, and E. Vanden-Eijnden. Adaptive monte carlo augmented with normalizing flows. *Proc. Nat. Acad. Sci. (PNAS)*, 119(10):e2109420119, 2022.
- M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Roy. Stat. Soc. Ser. B*, 73(2):123–214, 2011.
- A. N. Gomez, M. Ren, R. Urtasun, and R. B. Grosse. The reversible residual network: Backpropagation without storing activations. In *Adv. in Neural Information Process. Systems (NeurIPS)*, 2017.
- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 2020.
- U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *J. Roy. Stat. Soc. Ser. B*, 1994.
- I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Adv. in Neural Information Process. Systems (NeurIPS)*, 2017.
- P. Hagemann and S. Neumayer. Stabilizing invertible neural networks using mixture models. *Inverse Problems*, 2021.
- M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Adv. in Neural Information Process. Systems (NeurIPS)*, 2017.
- J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Adv. in Neural Information Process. Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.
- M. Hoffman, P. Sountsov, J. V. Dillon, I. Langmore, D. Tran, and S. Vasudevan. Neutra-lizing bad geometry in Hamiltonian Monte Carlo using neural transport. *arXiv preprint arXiv:1903.03704*, 2019.
- C.-W. Huang, L. Dinh, and A. Courville. Augmented normalizing flows: Bridging the gap between generative flows and latent variable models. *arXiv preprint arXiv:2002.07101*, 2020.
- T. Issenhuth, U. Tanielian, J. Mary, and D. Picard. Optimal precision for GANs. *arXiv preprint arXiv:2207.10541*, 2022.
- P. Izmailov, P. Kirichenko, M. Finzi, and A. G. Wilson. Semi-supervised learning with normalizing flows. In *Proc. Int. Conf. Machine Learning (ICML)*. PMLR, 2020.
- A. Justel, D. Peña, and R. Zamar. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & probability letters*, 35(3):251–259, 1997.
- D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Adv. in Neural Information Process. Systems (NeurIPS)*, 2018.
- A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6):066138, June 2004. doi: 10.1103/PhysRevE.69.066138. Publisher: American Physical Society.
- A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (Canadian Institute for Advanced Research), 2010.
- A. Kumar, P. Sattigeri, and T. Fletcher. Semi-supervised learning with GANs: Manifold invariance with improved inference. In *Adv. in Neural Information Process. Systems (NeurIPS)*, 2017.
- M. Ledoux. Isoperimetry and Gaussian analysis. *Lectures on probability theory and statistics*, 1996.
- Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, December 2015.
- Y. Marzouk, T. Moseley, M. Parno, and A. Spantini. Sampling via measure transport: An introduction. *Handbook of uncertainty quantification*, 1:2, 2016.
- E. Milman and J. Neeman. The Gaussian double-bubble and multi-bubble conjectures. *Annals of Mathematics*, 2022.

- T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák. Neural importance sampling. *ACM Transactions on Graphics (ToG)*, 38(5):1–19, 2019.
- E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Do deep generative models know what they don’t know? *arXiv preprint arXiv:1810.09136*, 2018.
- D. Nielsen, P. Jaini, E. Hoogeboom, O. Winther, and M. Welling. SurVAE flows: Surjections to bridge the gap between VAEs and flows. In *Adv. in Neural Information Process. Systems (NeurIPS)*, 2020.
- F. Noé, S. Olsson, J. Köhler, and H. Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.
- B. Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003.
- G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In *Adv. in Neural Information Process. Systems (NeurIPS)*, 2017.
- G. Papamakarios, E. T. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learning Research*, 2021.
- G. G. Pires and M. A. Figueiredo. Variational mixture of normalizing flows. *arXiv preprint arXiv:2009.00585*, 2020.
- S. Rifai, G. Mesnil, P. Vincent, X. Muller, Y. Bengio, Y. Dauphin, and X. Glorot. Higher order contractive auto-encoder. In *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*. Springer, 2011.
- V. Runde, K. Ribet, and S. Axler. *A taste of topology*. Springer, 2005.
- S. Samsonov, E. Lagutin, M. Gabrié, A. Durmus, A. Naumov, and E. Moulines. Local-global MCMC kernels: the best of both worlds. In *Adv. in Neural Information Process. Systems (NeurIPS)*, volume 35, pages 5178–5193, 2022.
- V. Stimper, B. Schölkopf, and J. M. Hernández-Lobato. Resampling base distributions of normalizing flows. In *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, pages 4915–4936. PMLR, 2022.
- M. Vono, N. Dobigeon, and P. Chainais. High-dimensional Gaussian sampling: a review and a unifying approach based on a stochastic proximal point algorithm. *SIAM Rev.*, 64(1):3–56, 2022.
- H. Wu, J. Köhler, and F. Noé. Stochastic normalizing flows. In *Adv. in Neural Information Process. Systems (NeurIPS)*, 2020.
- T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, and M. Girolami. Langevin diffusions and the metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, 91:14–19, 2014.
- F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.