Metropolized Flow: from Invertible Flow to MCMC

Achille Thin [1] **Nikita Kotelevskii** [2] **Alain Durmus** [3] **Maxim Panov** [2] **Eric Moulines** [1]

Abstract

In this contribution, we propose a new computationally efficient method *MetFlow* to combine Variational Inference (VI) with MCMC. In this approach, the standard mean-field variational distribution is enriched with MCMC transitions with proposals obtained using Normalizing Flows. The marginal distribution produced by such algorithm is a mixture of flow-based distributions, thus drastically increasing the expressivity of the variational family. Extensive numerical experiments show clear computational and performance improvements over state-of-the-art methods.

1. Introduction

One of the biggest computational challenge these days in machine learning and computational statistics is to sample from a complex distribution known up to a multiplicative constant. Indeed, this problem naturally appears in Bayesian inference (Robert, 2007) or for generative models like Variational AutoEncoders (VAE) or energy-based models (Kingma & Welling, 2013). Popular methods to address this problem are Markov Chain Monte Carlo (MCMC) algorithms (Brooks et al., 2011) and Variational Inference (VI) (Wainwright et al., 2008; Blei, 2017).

Starting from a parameterized family of distributions $\mathcal{Q} = \{q_\phi \colon \phi \in \Phi \subset \mathbb{R}^q\}$, VI approximates the intractable distribution with density π on $\mathbb{R}^D$ by maximizing the evidence lower bound (ELBO) defined by

$$\mathcal{L}(\phi) = \int \log\big(\tilde{\pi}(z)/q_\phi(z)\big) q_\phi(z) \mathrm{d}z \;, \tag{1}$$

using an unnormalized version $\tilde{\pi}$ of π, *i.e.* $\pi = \tilde{\pi}/\mathrm{C}_\pi$ setting $\mathrm{C}_\pi = \int_{\mathbb{R}^D} \tilde{\pi}(z)\mathrm{d}z$. Indeed, this approach consists in minimizing $\phi \mapsto \mathrm{KL}(q_\phi|\pi)$ since $\mathcal{L}(\phi) = \log(\mathrm{C}_\pi) - \mathrm{KL}(q_\phi|\pi)$.

[1]CMAP, Ecole Polytechnique, Universite Paris-Saclay, 91128 Palaiseau, France [2]CDISE, Skolkovo Institute of Science and Technology, Moscow, Russia [3]Ecole Normale Supérieure Paris-Saclay, Cachan. Correspondence to: Achille Thin <achille.thin@polytechnique.edu>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML).

The design of the family $\mathcal{Q}$ of variational distributions has a significant impact on the overall performance.

Recently, it has been suggested to enrich the traditional mean field variational approximation by combining them with invertible mappings with additional trainable parameters. A popular implementation of this principle is the Normalizing Flows (NFs) approach (Dinh et al., 2016; Rezende & Mohamed, 2015; Kingma et al., 2016) in which a mean-field variational distribution is deterministically transformed through a fixed-length sequence of parameterized invertible mappings.

The drawback of variational methods is that they only allow the target distribution to be approximated by a parametric family of distributions. On the contrary, MCMC are generic methods which have theoretical guarantees (Robert & Casella, 2013). The basic idea behind MCMC is to design a Markov chain $(Z_k)_{k\in\mathbb{N}}$ whose stationary distribution is π. Under mild assumptions, the distribution of Z_K converges to the target π as K goes to infinity. Yet, this convergence is in most cases slow and therefore this class of methods can be prohibitively computationally expensive. The idea to "bridge the gap" between MCMC and VI was first considered in (Salimans et al., 2015) and has later been pursued in several works; see (Wolf et al., 2016), (Hoffman et al., 2019) and (Caterini et al., 2018) and the references therein. We follow this line of research and design a new methodology combining VI and MCMC. More specifically, our contributions are as follows.

(1) We derive a new computationally tractable ELBO which allows us to use a large class of MCMC methods for the design of novel variational families.
(2) We then construct a new Metropolis-Hastings algorithm *MetFlow* which can take advantage of the full computational potential of Normalizing Flows. Contrary to classical variational inference using NFs, MetFlow guarantees that the target distribution is invariant.
(3) We combine MetFlow and our new ELBO to obtain a rich variational family which can be efficiently optimized.
(4) Finally, we present several numerical illustrations to show that our approach allows us to meaningfully trade-off between the approximation of the target distribution and computations, improving over state-of-the-art methods.

2. A New Combination Between VI and MCMC

Basics of Metropolis-Hastings The Metropolis Hastings (MH) algorithm to sample a density π w.r.t. the Lebesgue measure on $\mathbb{R}^D$ defines a Markov chain $(Z_k)_{k\in\mathbb{N}}$ with stationary distribution π as follows. Let $(U_k)_{k\in\mathbb{N}^*}$ be a sequence of i.i.d. random variables valued in $(\mathsf{U},\mathcal{U})$, with density h w.r.t. to a measure μ_{U}, and $T_\phi\colon \mathbb{R}^D\times\mathsf{U}\to\mathbb{R}^D$ be a function parameterized by $\phi\in\Phi$[1]. (U_k) is referred to as the *innovation noise* and T_ϕ as the *proposal mapping*. Conditionally to the current state $Z_k\in\mathbb{R}^D$, $k\in\mathbb{N}$, a proposal $Y_{k+1}=T_\phi(Z_k,U_{k+1})$ is sampled. Then, $Z_{k+1}=Y_{k+1}$ with probability $\alpha_\phi^{\mathsf{MH}}\big(Z_k,T_\phi(Z_k,U_{k+1})\big)$ and $Z_{k+1}=Z_k$ otherwise. The acceptance ratio $\alpha_\phi^{\mathsf{MH}}\colon\mathbb{R}^{2D}\to[0,1]$ is designed so that the resulting Markov kernel, denoted by $M_{\phi,h}$, is reversible w.r.t. π.

With this notation, $M_{\phi,h}$ can be written, for $z\in\mathbb{R}^D$, $\mathsf{A}\in\mathcal{B}(\mathbb{R}^D)$, as $M_{\phi,h}(z,\mathsf{A})=\int_{\mathsf{U}}h(u)Q_\phi\big((z,u),\mathsf{A}\big)\mu_{\mathsf{U}}(\mathrm{d}u)$ where for any $z\in\mathbb{R}^D$, $u\in\mathsf{U}$, $\mathsf{A}\in\mathcal{B}(\mathbb{R}^D)$,

$$Q_\phi((z,u),\mathsf{A})=\alpha_\phi(z,u)\delta_{T_\phi(z,u)}(\mathsf{A}) + \big\{1-\alpha_\phi(z,u)\big\}\delta_z(\mathsf{A})\ . \quad (2)$$

In this definition, δ_z stands for the Dirac measure at z and $\{\alpha_\phi\colon\mathbb{R}^D\times\mathsf{U}\to[0,1]\ ,\ \phi\in\Phi\}$ is a family of acceptance functions related to the MH acceptance probabilities by $\alpha_\phi(z,u)=\alpha_\phi^{\mathsf{MH}}\big(z,T_\phi(z,u)\big)$.

Variational Inference Meets Metropolis-Hastings Let $K\in\mathbb{N}^*$, $\{\xi_\phi^0\colon\phi\in\Phi\}$ on $\mathbb{R}^D$ be a parametric family of distributions and $\{h_i\}_{i=1}^K$ be density functions w.r.t μ_{U}. Consider now the following variational family

$$\mathcal{Q}=\{\xi_\phi^K=\xi_\phi^0 M_{\phi,h_1}\cdots M_{\phi,h_K}\colon\phi\in\Phi\}\ , \quad (3)$$

obtained by iteratively applying to the initial distribution ξ_ϕ^0 the Markov kernels $(M_{\phi,h_i})_{i=1}^K$.

For any $\phi\in\Phi$, $u\in\mathsf{U}$ and $z\in\mathbb{R}^D$, denote by $T_{\phi,u}(z)=T_\phi(z,u)$, $\alpha_{\phi,u}(z)=\alpha_\phi(z,u)$ and similarly for any $\mathsf{A}\in\mathcal{B}(\mathbb{R}^D)$, $Q_{\phi,u}(z,\mathsf{A})=Q_\phi\big((z,u),\mathsf{A}\big)$.

The key assumption in this section is that for any $\phi\in\Phi$ and $u\in\mathsf{U}$, $T_{\phi,u}$ is a C^1 diffeomorphism. This property is satisfied under mild condition on the proposal mapping. Indeed, one of our main result is that if ξ_ϕ^0 has density w.r.t. the Lebesgue measure, then ξ_ϕ^K as well. This is the crux of the construction of our new ELBO in order to use $\mathcal{Q}$ as a variational family.

For a $\mathrm{C}^1(\mathbb{R}^D,\mathbb{R}^D)$ diffeomorphism ψ, define by $J_\psi(z)$ the absolute value of the Jacobian determinant at $z\in\mathbb{R}^D$. For a family $\{T_i\}_{i=1}^K$ of mappings on $\mathbb{R}^D$ and $1\le i\le k<K$, define $\bigcirc_{j=i}^k T_j=T_i\circ\cdots\circ T_k$, for a sequence of vectors $(x_i)_{i=1}^K$ note $\mathbf{x}_K=(x_i)_{i=1}^K$ and for a sequence $\{h_i\}_{i=1}^K$ of innovation noise densities w.r.t. μ_{U}, define $\mathbf{h}_K(\mathbf{u}_K)=\prod_{i=1}^K h_i(u_i)$. Finally, set $\alpha_{\phi,u}^1(z)=\alpha_{\phi,u}(z)$ and $\alpha_{\phi,u}^0(z)=1-\alpha_{\phi,u}(z)$.

Proposition 1. *Assume that for any $(u,\phi)\in\mathsf{U}\times\Phi$, $T_{\phi,u}$ is a C^1 diffeomorphism and ξ_ϕ^0 admits a density m_ϕ^0 w.r.t. the Lebesgue measure. For any $\{u_i\}_{i=1}^K\in\mathsf{U}^K$, the distribution $\xi_\phi^K(\cdot|\mathbf{u}_K)=\xi_\phi^0Q_{\phi,u_1}\cdots Q_{\phi,u_K}$ has a density m_ϕ^K given by $m_\phi^K(z|\mathbf{u}_K)=\sum_{\mathbf{a}_K\in\{0,1\}^K}m_\phi^K(z,\mathbf{a}_K|\mathbf{u}_K)$ where*

$$m_\phi^K(z,\mathbf{a}_K|\mathbf{u}_K)=\prod_{i=1}^K\alpha_{\phi,u_i}^{a_i}\big(\bigcirc_{j=i}^K T_{\phi,u_j}^{-a_j}(z)\big) \times m_\phi^0\big(\bigcirc_{j=1}^K T_{\phi,u_j}^{-a_j}(z)\big)J_{\bigcirc_{j=1}^K T_{\phi,u_j}^{-a_j}}(z)\ . \quad (4)$$

In particular, for a sequence $\{h_i\}_{i=1}^K$ of innovation noise densities, ξ_ϕ^K (3) has a density w.r.t. the Lebesgue measure, explicitly given, for any $z\in\mathbb{R}^D$, by $m_\phi^K(z)=\int_{\mathsf{U}^K}\Big\{m_\phi^K(z|\mathbf{u}_K)\mathbf{h}_K(\mathbf{u}_K)\Big\}\,\mathrm{d}\mu_{\mathsf{U}}^{\otimes K}(\mathbf{u}_K)$.

We can now apply the VI approach the family $\mathcal{Q}$ defined in (3). As $m_\phi^K(z)$ is intractable directly, we define $\mathcal{L}_{\mathrm{aux}}(\phi)$

$$\mathcal{L}_{\mathrm{aux}}(\phi)=\sum_{\mathbf{a}_K\in\{0,1\}^K}\int\mathbf{h}_K(\mathbf{u}_K)m_\phi^K(z_K,\mathbf{a}_K|\mathbf{u}_K) \times\log\left(\frac{2^{-K}\tilde{\pi}(z_K)}{m_\phi^K(z_K,\mathbf{a}_K|\mathbf{u}_K)}\right)\mathrm{d}z_K\mathrm{d}\mu_{\mathsf{U}}^{\otimes K}(\mathbf{u}_K)\ . \quad (5)$$

Note that using Jensen's inequality w.r.t. the density $(\mathbf{u}_K,\mathbf{a}_K)\mapsto\mathbf{h}_K(\mathbf{u}_K)m_\phi^K(\mathbf{a}_K|z_K,\mathbf{u}_K)$, we get $\mathcal{L}_{\mathrm{aux}}(\phi)\le\mathcal{L}(\phi)$. The ELBO $\mathcal{L}_{\mathrm{aux}}$ can be optimized w.r.t. ϕ, typically by stochastic gradient methods, which requires an unbiased estimator of the gradient $\nabla\mathcal{L}_{\mathrm{aux}}(\phi)$. Such estimators are amenable to the reparameterization trick (Rezende et al., 2014).

3. MetFlow: MCMC and Normalizing Flows

We present in what follows a new class of MCMC methods, Metropolized Flows (MetFlow), which can be cast in the framework introduced in the previous section and for which the proposal mappings are Normalizing Flows (NF). This class of methods then naturally defines a variational family $\mathcal{Q}$ of the form (3) for which the parameter ϕ is optimized using (5). Our objective is to capitalize on the flexibility of NFs to represent distributions, while keeping the theoretical guarantees of MCMC.

Consider a flow $T_\phi\colon\mathbb{R}^D\times\mathsf{U}\to\mathbb{R}^D$ parametrized by $\phi\in\Phi$. It is assumed that for any $u\in\mathsf{U}$, $T_{\phi,u}\colon z\mapsto T_\phi(z,u)$ is a C^1 diffeomorphism. Set $\mathsf{V}=\{-1,1\}$. For any $u\in\mathsf{U}$, consider the involution $\mathring{T}_{\phi,u}$ on $\mathbb{R}^D\times\mathsf{V}$, *i.e.* $\mathring{T}_{\phi,u}\circ\mathring{T}_{\phi,u}=$

[1] In this work, ϕ collectively denotes the parameters used in the proposal distribution

Id, defined for $z \in \mathbb{R}^D$, $v \in \{-1, 1\}$ by

$$\mathring{T}_{\phi,u}(z, v) = (T^v_{\phi,u}(z), -v) \; . \tag{6}$$

The variable v is called the direction. If $v = 1$ (respectively $v = -1$), the "forward"(resp. "backward") flow $T_{\phi,u}$ (resp. $T^{-1}_{\phi,u}$) is used. For any $z \in \mathbb{R}^D$, $v \in \{-1, 1\}$, $\mathsf{A} \in \mathcal{B}(\mathbb{R}^D)$, $\mathsf{B} \subset \mathsf{V}$, we define the kernel

$$\begin{aligned} R_{\phi,u}\big((z, v), \mathsf{A} \times \mathsf{B}\big) = \mathring{\alpha}_{\phi,u}(z, v)\delta_{T^v_{\phi,u}(z)}(\mathsf{A}) \otimes \delta_{-v}(\mathsf{B}) \\ + \{1 - \mathring{\alpha}_{\phi,u}(z, v)\}\delta_z(\mathsf{A}) \otimes \delta_v(\mathsf{B}) \; , \end{aligned} \tag{7}$$

where $\mathring{\alpha}_{\phi,u} \colon \mathbb{R}^D \times \mathsf{V} \to [0, 1]$ is the acceptance function.

Proposition 2. *Let ν be a distribution on V, and $(u, \phi) \in \mathsf{U} \times \Phi$. Assume that $\mathring{\alpha}_{\phi,u} \colon \mathbb{R}^D \times \mathsf{V} \to [0, 1]$ satisfies for any $(z, v) \in \mathbb{R}^D \times \mathsf{V}$,*

$$\begin{aligned} &\mathring{\alpha}_{\phi,u}(z, v)\pi(z)\nu(v) \\ &\quad = \mathring{\alpha}_{\phi,u}\big(\mathring{T}_\phi(z, v)\big)\pi\big(T^v_{\phi,u}(z)\big)\nu(-v)J_{T^v_{\phi,u}}(z) \; . \end{aligned} \tag{8}$$

Then for any $(u, \phi) \in \mathsf{U} \times \Phi$, $R_{\phi,u}$ defined by (7) is reversible with respect to $\pi \otimes \nu$. In particular, if for any $(z, v) \in \mathbb{R}^D \times \mathsf{V}$,

$$\mathring{\alpha}_{\phi,u}(z, v) = \varphi\left(\pi\big(T^v_{\phi,u}(z)\big)\nu(-v)J_{T^v_{\phi,u}}(z)/\pi(z)\nu(v)\right) \; ,$$

for $\varphi \colon \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}_+$, then (8) is satisfied if $\varphi(+\infty) = 1$ and for any $t \in \overline{\mathbb{R}}_+$, $t\varphi(1/t) = \varphi(t)$.

Remark 1. *The condition (8) on the acceptance ratio $\mathring{\alpha}_{\phi,u}$ has been reported in (Tierney, 1998, Section 2), and standard choices for $\mathring{\alpha}_{\phi,u}$ are the Metropolis-Hastings and Barker ratios which correspond to $\varphi \colon t \mapsto \min(1, t)$ and $t \mapsto t/(1 + t)$ respectively.*

If we define for $u \in \mathsf{U}$, $v \in \mathsf{V}$, $z \in \mathbb{R}^D$, $\mathsf{A} \in \mathcal{B}(\mathbb{R}^D)$,

$$\begin{aligned} &Q_{\phi,(u,v)}(z, \mathsf{A}) = R_{\phi,u}((z, v), \mathsf{A} \times \mathsf{V}) \\ &= \mathring{\alpha}_{\phi,u}(z, v)\delta_{T^v_{\phi,u}(z)}(\mathsf{A}) + \{1 - \mathring{\alpha}_{\phi,u}(z, v)\}\delta_z(\mathsf{A}) \; , \end{aligned} \tag{9}$$

we retrieve the framework defined in Section 2. In turn, from a distribution ν for the direction, the family $\{Q_{\phi,(u,v)} \colon (u, v) \in \mathsf{U} \times \mathsf{V}\}$ defines a MH kernel, given for $u \in \mathsf{U}$, $z \in \mathbb{R}^D$, $\mathsf{A} \in \mathcal{B}(\mathbb{R}^D)$ by

$$M_{\phi,u,\nu}(z, \mathsf{A}) = \nu(1)Q_{\phi,(u,1)}(z, \mathsf{A}) + \nu(-1)Q_{\phi,(u,-1)}(z, \mathsf{A}) \; . \tag{10}$$

The key result of this section is

Corollary 1. *For any $u \in \mathsf{U}$ and any distribution ν, the kernel $M_{\phi,u,\nu}$ is reversible w.r.t. π.*

As the reversibility is satisfied for any $\mathbf{u}_K \in \mathsf{U}^K$, we now focus on a fixed sequence $\mathbf{u}_K$ of proposal noise. This allows us to consider a setting close to NF where distribution pushforwards are deterministic. We thus write $m^K_{\phi,\mathbf{u}_K}(\cdot|\mathbf{v}_K) = m^K_\phi(\cdot|\mathbf{u}_K, \mathbf{v}_K)$.

The choice of the transformation T_ϕ is really flexible. Let $\{\mathsf{T}_{\phi,i}\}_{i=1}^K$ be a family of K diffeomorphisms on $\mathbb{R}^D$. A flow model based on $\{\mathsf{T}_{\phi,i}\}_{i=1}^K$ is defined as a composition $\mathsf{T}_{\phi,K} \circ \cdots \circ \mathsf{T}_{\phi,1}$ that pushes an initial distribution ξ^0_ϕ with density m^0_ϕ to a more complex target distribution ξ^K_ϕ with density m^K_ϕ, given for any $z \in \mathbb{R}^D$ by $m^K_\phi(z) = m^0\big(\bigcirc_{i=1}^K \mathsf{T}^{-1}_{\phi,i}(z)\big)J_{\bigcirc_{i=1}^K \mathsf{T}^{-1}_{\phi,i}}(z)$, see (Tabak & Turner, 2013; Rezende & Mohamed, 2015; Kobyzev et al., 2019; Papamakarios et al., 2019). We then construct a variational family $\mathcal{Q}$ of the form (2) based on (10) and the same deterministic sequence of diffeomorphisms. More precisely, a *MetFlow* model is obtained by applying successively the Markov kernels $M_{\phi,1,\nu}, \ldots, M_{\phi,K,\nu}$, written as, for $z \in \mathbb{R}^D$, $\mathsf{A} \in \mathcal{B}(\mathbb{R}^D)$, $i \in \{1, \ldots, K\}$:

$$\begin{aligned} M_{\phi,i,\nu}(z, \mathsf{A}) = \sum_{v \in \mathsf{V}} \nu(v)\mathring{\alpha}_{\phi,i}(z, v)\delta_{\mathsf{T}^v_{\phi,i}(z)}(\mathsf{A}) \\ + (1 - \sum_{v \in \mathsf{V}} \nu(v)\mathring{\alpha}_{\phi,i}(z, v))\delta_z(\mathsf{A}) \; . \end{aligned}$$

Each of those is reversible w.r.t. the stationary distribution π and thus leaves π invariant. In such a case, the resulting distribution ξ^K_ϕ is a mixture of the pushforward of ξ^0_ϕ by the flows $\{\mathsf{T}^{v_K a_K}_{\phi,K} \circ \cdots \circ \mathsf{T}^{v_1 a_1}_{\phi,1}, \, \mathbf{v}_K \in \mathsf{V}^K, \mathbf{a}_K \in \{0, 1\}^K\}$. The parameters ϕ of the flows $\{\mathsf{T}_{\phi,i}\}_{i=1}^K$ are optimized by maximizing the ELBO defined by (5) in which $m^K_{\phi,\mathbf{u}_K}$ is substituted by $m^K_{\phi,1:K}$ with $T_{\phi,u_i} \leftarrow \mathsf{T}_{\phi,i}$.

Among the different flow models which have been considered recently in the literature (Papamakarios et al., 2019), we chose Real-Valued Non-Volume Preserving (RNVP) flows (Dinh et al., 2016) because they are easy to compute and invert. Other implementations are left to future work (Ho et al., 2019; Kingma et al., 2016; Huang et al., 2018; Cao et al., 2019; Wehenkel & Louppe, 2019).l

4. Experiments

We illustrate in this section the computational benefits of our new VI approach using a MetFlow model using RNVP flows. We present examples of sampling from complex synthetic distributions which are often used to benchmark generative models, such as a mixture of highly separated Gaussians and other non-Gaussian 2D distributions. We also present posterior inference approximations and inpainting experiments on MNIST dataset, in the setting outlined by (Levy et al., 2017).

We consider two settings. In the *deterministic* setting, we use K different RNVP transforms $\{\mathsf{T}_{\phi,i}\}_{i=1}^K$, and the parameters for each individual transform $\mathsf{T}_{\phi,i}$ are different. In the *pseudo-randomized* setting, we define global transformation T_ϕ on $\mathbb{R}^D \times \mathsf{U}$ and set $\mathsf{T}_{\phi,i} = T_\phi(\cdot, u_i)$, where $\mathbf{u}_K$ are K independent draws from a standard normal distribution. In such case, the parameters are the same for the flows

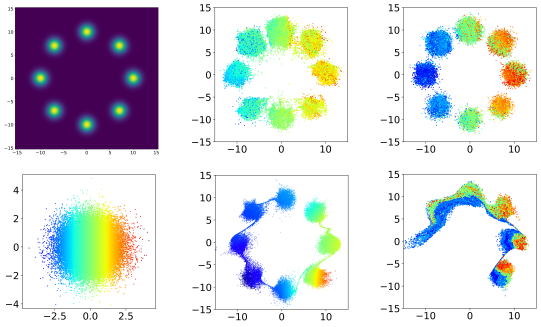

Figure 1. Sampling a mixture of 8 Gaussian distributions. Top row from left to right: Target distribution, MetFlow, MetFlow with 145 resampled innovation noise. Bottom row from left to right: Prior distribution, First run of RNVP, Second run of RNVP. MetFlow finds all the modes and improves with more iterations, while RNVP depend on a good initialization to find all the modes and fails to separate them correctly.

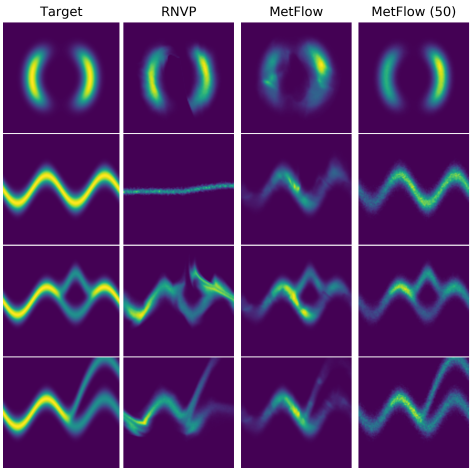

Figure 2. Density matching example (Rezende & Mohamed, 2015) and comparison between RNVP and MetFlow.

$\mathsf{T}_{\phi,i}$, only the innovation noise u_i differs. Typically, RNVP are encoded by neural networks. In the second setting, the network will thus take as input z and u stacked together.

In the second setting, once training has been completed and a fit $\hat{\phi}$ of the parameters has been obtained, we can sample additional noise innovations $(u_i)_{i=K+1}^{mK}$. We then consider the distribution given by $\xi_{\hat{\phi}}^{mK} = \xi_{\hat{\phi}}^{K} M_{\hat{\phi},u_{K+1},\nu}, \ldots, M_{\hat{\phi},u_{mK},\nu}$ where ν is typically the uniform on $\{-1,1\}$, as defined as in Section 3. mK corresponds to the length of the final Markov chain we consider. In practice, we have found that sampling additional noise innovations this way yields a more accurate approximation of the target, thanks to the asymptotic guarantees of MCMC.

4.1. Synthetic data. Examples of sampling.

Mixture of Gaussians The objective is to sample from a mixture of 8 Gaussians in dimension 2, starting from a standard normal prior distribution q^0, and compare MetFlow to RNVP. We are using an architecture of five RNVP flows ($K = 5$), each of which is parametrized by two three-layer fully-connected neural networks with LeakyRelu (0.01) activations. In this example, we consider the pseudo-randomized setting. The results for MetFlow and for RNVPs alone are shown on Figure 1. First, we observe that while our method successfully finds all modes of the target distribution, RNVP alone struggles to do the same. Our method is therefore able to approximate multimodal distributions with well separated modes. Here, the mixture structure of the distribution (with potentially $3^5 = 243$ modes) produced by MetFlow is very appropriate to such a problem. On the contrary, classical flows are unable to approximate well separated modes starting from a simple unimodal prior, without much surprise. In particular, mode dropping is a serious issue even in small dimension. Moreover, an other advantage of MetFlow in the pseudo randomized setting is to be able to iterate the learnt kernels which still preserve the target distribution. Iterating MetFlow kernels widens the gap between both approaches, significantly improving the accuracy of our approximation.

Non-Gaussian 2D Distributions In a second experiment, we sample the non-Gaussian 2D distributions proposed in (Rezende & Mohamed, 2015). Figure 2 illustrates the performance of MetFlow compared to RNVP. We are again using 5 RNVPs ($K = 5$) with the architecture described above, and use the pseudo-randomized setting for MetFlow. After only five steps, MetFlow already finds the correct form of the target distribution, while the simple RNVP fails on the more complex distributions. Moreover, iterating again MetFlow kernels allows us to approximate the target distribution with striking precision, after only 50 MCMC steps.

4.2. Deep Generative Models

Deep Generative Models (DGM), such as Deep Latent Gaussian Models (see Kingma & Welling (2013); Rezende et al. (2014)) have recently become very popular. The basic assumption in a DGM is that the observed data x is generated by sampling a latent vector z which is used as the input of a deep neural network. This network then outputs the parameters of a family of distributions (e.g., the canonical parameters of exponential family like Bernoulli or Gaussian distributions) from which the data are sampled. Given data generated by a DGM, a classical problem is to approach the conditional distribution $p(z \mid x)$ of the latent variables z given the observation x, using variational inference to construct an amortized approximation.

We consider the binarized MNIST handwritten digit dataset.

The generative model is as follows. The latent variable z is a $l = 64$ dimensional standard normal Gaussian. The observation $x = (x^j)_{j=1}^D$ is a vector of $D = 784$ bits. The bits $(x^j)_{j=1}^D$ are, given the latent variable z, conditionally independent Bernoulli distributed random variables with success probability $p_\theta(z)^j$ where $(p_\theta^j)_{j=1}^D$ is the output of a convolutional neural network. In this framework, p_θ is called the decoder. In the following, we show that our method provides a flexible and accurate variational approximation of the conditional distribution of the latent variable given the observation $p_\theta(z \mid x)$, outperforming mean-field and Normalizing Flows based approaches.

As we are focusing in this paper on the comparison of VI methods to approximate complex distributions and not on learning the Variational Auto Encoder itself, we have chosen to use a fixed decoder for both Normalizing Flows (here, Neural Autoregessive Flows) and MetFlow (with RNVP transforms). The decoder is obtained using state-of-the-art method. We can illustrate the expressivity of MetFlow in two different ways. We first fix L different samples. In this example, we take $L = 3$ images representing the digit "3". We are willing to approximate, for a given decoder p_θ, the posterior distribution $p_\theta(z|(x_i)_{i=1}^L)$. We show in Figure 3 the decoded samples corresponding to the following variational approximations of $p_\theta(\cdot|(x_i)_{i=1}^L)$: (i) a NAF trained from the decoder to approximate $p_\theta(\cdot|(x_i)_{i=1}^L)$ and (ii) *MetFlow* in the deterministic setting with $K = 5$ RNVP flows.

Figure 3 shows that the samples generated from (i) collapse essentially to one mode corresponding to the first digit. On the contrary, MetFlow is able to capture the three different modes of the posterior and generates much more variability in the decoded samples. We now consider the in-painting set-up introduced in (Levy et al., 2017, Section 5.2.2). Formally, we in-paint the top of an image using Block Gibbs sampling. Given an image x, we denote x^t, x^b the top and the bottom half pixels. Starting from an image $\mathbf{x_0}$, we sample at each step $z_t \sim p_\theta(z \mid x_t)$ and then $\tilde{x} \sim p_\theta(x \mid z_t)$. We the set $x_{t+1} = (\tilde{x}^t, x_0^b)$. We give the output of this process when sampling from the mean-field approximation of the posterior only, the mean-field pushed by a NAF, or using our method. The result for the experiment can be seen on Figure 4.

We can see that MetFlow mixes easily between different modes, and produces sharp images. We recognize furthermore different digits (3,5,9). It is clear from the middle plot that the mean-field approximation is not able to capture the complexity of the distribution $p_\theta(z \mid x)$. Finally, the NAF improves the quality of the samples but does not compare to MetFlow in terms of mixing.

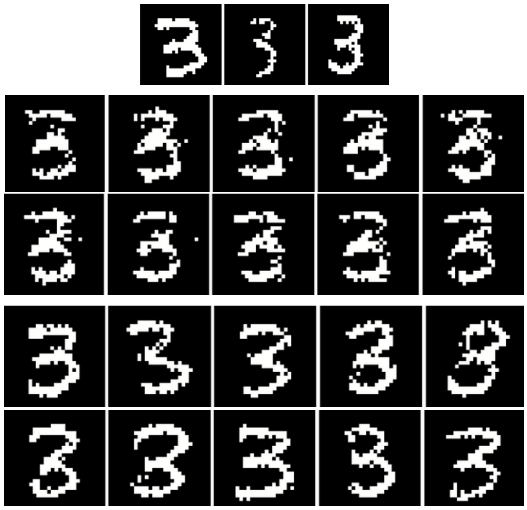

Figure 3. Mixture of '3' digits. Top: Fixed digits, Middle: NAF samples, Bottom: MetFlow samples. Compared to NAF, MetFlow is capable to mix better between these modes, while NAF seems to collapse.

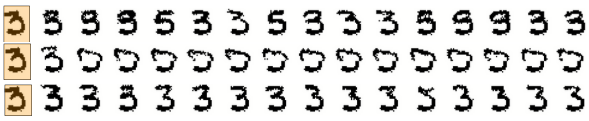

Figure 4. Top line: Mean-Field approximation and MetFlow, Middle line: Mean-Field approximation, Bottom line: Mean-Field Approximation and NAF. Orange samples on the left represent the initialization image. We observe that MetFlow easily mixes between the modes while other methods are stuck in one mode.

References

Blei, D., K. A. M. J. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, Feb 2017. ISSN 1537-274X. doi: 10.1080/01621459.2017.1285773.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

Cao, N. D., Titov, I., and Aziz, W. Block neural autoregressive flow. In *UAI*, 2019.

Caterini, A. L., Doucet, A., and Sejdinovic, D. Hamiltonian variational auto-encoder. In *Advances in Neural Information Processing Systems*, pp. 8167–8177, 2018.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. *arXiv preprint arXiv:1605.08803*, 2016.

Ho, J., Chen, X., Srinivas, A., Duan, Y., and Abbeel, P. Flow++: Improving flow-based generative models with variational dequantization and architecture design. *arXiv preprint arXiv:1902.00275*, 2019.

Hoffman, M., Sountsov, P., Dillon, J. V., Langmore, I., Tran, D., and Vasudevan, S. Neutra-lizing bad geometry in Hamiltonian Monte Carlo using neural transport. *arXiv preprint arXiv:1903.03704*, 2019.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. Neural autoregressive flows. *arXiv preprint arXiv:1804.00779*, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pp. 4743–4751, 2016.

Kobyzev, I., Prince, S., and Brubaker, M. A. Normalizing flows: Introduction and ideas. *arXiv preprint arXiv:1908.09257*, 2019.

Levy, D., Hoffman, M. D., and Sohl-Dickstein, J. Generalizing Hamiltonian Monte Carlo with neural networks. *arXiv preprint arXiv:1711.09268*, 2017.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Robert, C. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

Robert, C. and Casella, G. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Salimans, T., Kingma, D., and Welling, M. Markov chain Monte Carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pp. 1218–1226, 2015.

Tabak, E. G. and Turner, C. V. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

Tierney, L. A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9, 02 1998.

Wainwright, M. J., Jordan, M. I., et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Wehenkel, A. and Louppe, G. Unconstrained monotonic neural networks. In *Advances in Neural Information Processing Systems*, pp. 1543–1553, 2019.

Wolf, C., Karl, M., and van der Smagt, P. Variational inference with Hamiltonian Monte Carlo. *arXiv preprint arXiv:1609.08203*, 2016.