Sliced-Wasserstein normalizing flows: beyond maximum likelihood training

Florentin Coeurdoux[1], Nicolas Dobigeon[1,2] Pierre Chainais[3*]

1- University of Toulouse, IRIT/INP-ENSEEIHT, F-31071 Toulouse, France

2- Institut Universitaire de France (IUF), France

3- Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

Abstract. Despite their advantages, normalizing flows generally suffer from several shortcomings including their tendency to generate unrealistic data (e.g., images) and their failing to detect out-of-distribution data. One reason for these deficiencies lies in the training strategy which traditionally exploits a maximum likelihood principle only. This paper proposes a new training paradigm based on a hybrid objective function combining the maximum likelihood principle (MLE) and a sliced-Wasserstein distance. Results obtained on synthetic toy examples and real image data sets show better generative abilities in terms of both likelihood and visual aspects of the generated samples. Reciprocally, the proposed approach leads to a lower likelihood of out-of-distribution data, demonstrating a greater data fidelity of the resulting flows.

1 Introduction

Approximating probability distributions thanks to normalizing flows (NFs) has proven to be a powerful approach to accurately represent the underlying processes at the origin of collected data. NFs are designed to provide a tractable approximation of the data-generating density p_X by transforming a base normal distribution through a series of bijective transformations [1]. The usual approach to train such architectures relies on the principle of maximum likelihood estimation (MLE), i.e., maximizing the joint density of the observed data with respect to (w.r.t.) the parameters of the network. The constraint of relying on a parametric family of distributions may however raise crucial issues. Indeed, the optimality of MLE holds only when there is no model misspecification, i.e., the true data distribution p_X belongs to the family that can be represented by the optimized model. In practice, it is difficult to ensure a priori that the chosen family of functions is able to accurately model the targeted distribution. Hence the choice of the learning objective becomes largely an essential but most often empirical question. Moreover, training NF explicitly uses a Gaussian likelihood, i.e, a function of the two first moments. Hence, optimizing a Gaussian likelihood function leaves any higher order moments completely free. A more refined way of fully characterizing the targeted distribution would be to match all the other higher order moments as well. Clearly, using a statistical distance

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

between distributions during the training process would shift the optimization task from a nonlinear regression problem w.r.t. the likelihood parameters to a more relevant problem of looking for the best matching between the generated distribution and the targeted one. Grover *et al.* [2] proposed a hybrid objective that bridges implicit and prescribed learning by combining MLE and adversarial training using a GAN. The hybrid objective has a balancing effect between perceptually good-looking samples and an accurate density estimation of the inputs. The authors also demonstrate that this hybrid objective has a regularizing effect, which permits the model to outperform MLE as well as adversarial learning. However the choice of using an adversarial architecture is accompanied by the well-documented drawbacks of GANs. An adversarial architecture requires the training of an additional discriminator which is notoriously unstable, can lead to mode collapse [3] and can produce overconfident predictions from out-of-distribution (OoD) inputs [4].

To overcome the issues mentioned above, this paper introduces a novel hybrid loss function to be used to train NF. As suggested above, in addition to the conventional MLE-based term, the proposed hybrid training loss also incorporates a term to measure the discrepancy between the generated and the targeted distribution. This term derives from the Sliced-Wasserstein distance (SW) [5] between the true data distribution and the generated samples. Experimental results show that augmenting the MLE objective with this term consistently achieves higher likelihood as well as better quality of the generated samples. It also demonstrates better OoD detection capabilities compared to classical training of flow-based models. Section 2 introduces the proposed method referred to as sliced-Wasserstein NF (SW-NF) and its hybrid learning objective. Section 3 illustrates its performances on numerical experiments. Conclusions and prospects are reported in Section 4.

2 Sliced-Wasserstein flows

2.1 Normalizing flows

NFs are a flexible class of deep generative networks that learn a change of variable between two probability distributions p_X and p_Z through an invertible transformation $f_\theta : X \mapsto Z = f_\theta(X)$ parametrized by θ [1]. In general, p_X is only known through samples $x = \{x_n\}_{n=1}^N$ and, for tractability purpose, p_Z is chosen as a centered normal distribution with unit variance. The parameters θ defining the operator f_θ are then adjusted according the MLE principle and exploiting the change of variable

$$p_X(x) = p_Z\left(f_\theta(x)\right)\left|\det J_{f_\theta^{-1}}\right| \quad \text{with} \quad J_{f_\theta^{-1}} = \frac{\partial f_\theta^{-1}}{\partial x} \tag{1}$$

In other words, the network is trained by minimizing the negative log-likelihood (NLL) or equivalently the loss function denoted by $\mathcal{L}_{\text{MLE}}(x;\theta) = -\log(p_X(x))$. Without loss of generality, this work focuses on NFs based on affine coupling layers. Examples of such flows include RealNVP [6], Glow [7] among others.

2.2 Sliced-Wasserstein distance

In recent years, Wasserstein distance, which is intimately related to the theory of optimal transport (OT), has received a considerable attention from the machine learning (ML) community because of its theoretical properties when comparing distribution. However, it suffers from strong computational and statistical limitations, which have severely hindered its effective use in problems in high dimensions. Several workarounds have been proposed to alleviate these issues and to enable the use of OT in ML applications. In particular, the Sliced-Wasserstein (SW) distance is an alternative OT metric [5]. It has been increasingly popular since it benefits from a significantly reduced computational cost over the Wasserstein distance, especially on large-scale problems. In a nutshell, the SW distance compares high-dimensional distributions by comparing their projected 1d-distributions for which the computation of the Wasserstein distance is closed-form. According to a Monte Carlo principle the SW distance between two distributions p_X and p_Z empirically represented by two sets of samples x and z, respectively, can be approximated by *i)* drawing a large set of vectors $u_1, \ldots, u_J$ uniformly distributed over the unit sphere then *ii)* averaging the true 1d-Wasserstein distances between the slices of the two distributions along directions u_i. It will be denoted as $\mathcal{L}_{\mathrm{SW}}(x, z)$ in what follows. Its formulation through its projections onto the unit sphere is well adapted when samples are vectors. Introduced by Nguyen et *al.* [8], the so-called convolution SW (CSW) generalizes SW to images using a series of convolutions in the spirit of a multiresolution approach. We denote $\mathcal{L}_{\mathrm{CSW}}(x, z)$ the corresponding distance measure.

2.3 Hybrid objective function

The proposed SW-NF method builds on a NF neural architecture f_θ to target a normal latent distribution p_Z so that the likelihood of the observed data p_X is well-defined and tractable for exact evaluation and MLE training. Departing from conventional strategies deployed to train NFs, this work proposes to derive a hybrid objective function that binds the likelihood of a prescribed model to high order moment matching. To this aim the conventional MLE-based objective is augmented with an additional term measuring the discrepancy between the respective distributions of the original data $x \sim P_X$ and the generated data $\tilde{x} = f_\theta^{-1}(z)$. Note that the likelihood loss is prescribed on the latent space while the SW-based distance between the generated and target distributions can be prescribed over the data space. Thus the proposed hybrid objective is a combination of reconstruction and feature losses defined as

$$\mathcal{L}(x, z; \theta) = \mathcal{L}_{\star \mathrm{W}}(x, f_\theta^{-1}(z)) + \alpha \mathcal{L}_{\mathrm{MLE}}(x; \theta) \tag{2}$$

where α is a hyperparameter balancing the two terms and $\mathcal{L}_{\star \mathrm{W}}$ refers to either $\mathcal{L}_{\mathrm{SW}}$ for vector data sets or to $\mathcal{L}_{\mathrm{CSW}}$ for image inputs, respectively. It is worth noting that the new objective function can be interpreted as a regularized counterpart of the change of variable on the data space. Moreover it has the great advantage of not depending on an auxiliary network as in [2]. Note

Objective	NLL	SW	$\|\kappa_3\|_2^2$	$\|\kappa_4\|_2^2$
MLE	0.52	0.0033	0.2233	5.6124
SW	1.78	**0.0007**	**0.0026**	**0.1822**
SW-Flow	**0.41**	0.0008	0.0501	0.2259
Flow-GAN [2]	0.51	1.23	0.4756	7.7725

Table 1: Circle data set: assessment of goodness-of-fit (for all metrics, the lower the better).

Objective	Inception	NLL (bits/dim)	CSW	$\|\kappa_3\|_2^2$	$\|\kappa_4\|_2^2$
MLE	2.42	3.54	1514.26	64.94	2462.37
SW	1.28	9.81	1190.11	13.13	**598.54**
SW-Flow	3.04	**3.19**	**1014.26**	**6.24**	656.37
Flow-GAN	**3.21**	4.21	1621.78	72.3	3079.12

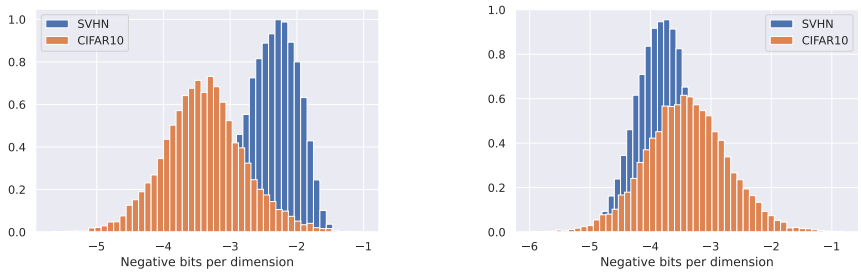
Table 2: CIFAR-10 data set: assessment of goodness-of-fit (for inception score, the higher the better; for all other metrics, the lower the better).

that the SW-based discrepancy measure between the generative model and the data distributions can also be prescribed over the latent space by replacing the SW-based term in (2) by $L_{\star \mathrm{W}}(z, f_\theta(x))$.

3 Numerical experiments

This section assesses the versatility and the accuracy of proposed SW-NF method through numerical experiments. First, experiments conducted on the Circle data set from scikit-learn are presented to provide some insights about key ingredients of the proposed approach. Then the performance of SW-NF is illustrated through more realistic experimental settings exploiting the CIFAR-10 and SVHN image data sets. It is compared to the alternative training strategies which consist in relying on the sole MLE or SW terms in (2) and to the Flow-GAN method which hybridizes MLE and GAN losses [2]. For all results reported below, the stochastic gradient descent is implemented in Pytorch, with the Adam optimizer, a learning rate of 10^{-4} and a batch size of 4096 or 8192 samples. When dealing with the toy example, the NF implementing the unknown mapping f is chosen as a RealNVP [6] and the conventional SW distance is chosen for hybridization. When dealing with the image-driven experiments, the network architecture is Glow [7] and the CSW is considered as a statistical distance. The proposed learning strategy is compared with conventional method from two task-driven perspectives, namely goodness-of-fit and OoD detection.

Goodness-of-fit: We first study the goodness-of-fit of the targeted latent space through the learned inverse transform. This evaluation is conducted by evaluating not only the NLL but also the (C)SW distance and the 3rd- and 4th-order cumulants κ_3 and κ_4 which are expected to be equal to zero for the prescribed normal distribution p_Z. Table 1 and Table 2 report the results reached by the



(a) MLE training (b) Hybrid training

Fig. 1: Likelihood histogram for 1000 CIFAR-10 images (orange) and 1000 SVHN images (blue) prescribed by a CSW-NF trained on CIFAR-10 images.

proposed SW-NF approach for the circle-shaped and the CIFAR-10 data sets, respectively. Table 2 also gives the inception score for visual inspection of the images. Interestingly, the proposed SW-NF method provides significantly better NLL scores than the sole MLE-based learning strategy. For the two data sets, it also leads to competitive results in terms of (C)SW and normality features, reaching scores close to SW-based learning. In addition, when considering the CIFAR-10 images, it leads to higher inception score, thus suggesting higher visual quality of the generated samples.

Method	Moons	Blobs
MLE	0.37	0.54
SW	**0.61**	0.70
SW-NF	0.53	**0.73**

Table 3: Circle data set: performance of OoD in term of AUROC (the closer to 1 the better).

Out-of-distribution detection: The second set of experiments assesses the ability of detecting OoD data. Table 3 reports the area under the receiving operator caracteristics (AUROC) for moon-shaped and blob-shaped data sets given a model trained on circles. In the absence of SW term in the training loss, the model constantly shows lower ability to discriminate OoD data from the training distribution data. In the context of CIFAR-10 data sets, the experimental setup of [9] has been considered. Glow-based NFs have been trained on CIFAR-10 and we monitor the prescribed likelihood for both CIFAR-10 and SVHN images. Fig. 1 (a) shows the results obtained by a sole MLE-based training: this model predicts a higher likelihood of OoD data. Fig. 1 (b) depicts similar plots when considering the proposed hybrid loss: it leads to lower likelihood to OoD data coming from the SVHN data set.

4 Conclusion

This paper introduces a new paradigm to train normalizing flow. It consists in augmenting the loss term derived from the conventional maximum likelihood principle with a discrepancy measure between the generated and targeted distributions. The resulting hybrid loss function thus combines a Gaussian likelihood with a (convolutional) sliced-Wasserstein distance between distributions. Numerical experiments show the better performance of the proposed hybrid training procedure in terms of perceptual as well as statistical quantitative metrics. On top of that, one observes a better robustness of the out-of-distribution behavior. This works consists of a step towards the design of more powerful NF implemented as true generative models, beyond their simple use as nonlinear regressors structurally imposed by a conventional MLE training.

References

[1] G Papamakarios, E Nalisnick, D J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[2] A. Grover, M. Dhar, and S. Ermon. Flow-gan: Combining maximum likelihood and adversarial learning in generative models. In *Proc. AAAI Conference on Artificial Intelligence*, 2018.

[3] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Proc. Advances in Neural Information Processing Systems*, volume 29, 2016.

[4] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proc. International Conference on Learning Representations*, 2017.

[5] J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Proc. Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer Berlin Heidelberg, 2012.

[6] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using Real NVP, 2017.

[7] D. P Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Proc. Advances in Neural Information Processing Systems*, 2018.

[8] K. Nguyen and N. Ho. Revisiting sliced wasserstein on images: From vectorization to convolution, 2022.

[9] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Do deep generative models know what they don't know? In *Proc. International Conference on Learning Representations*, 2019.