Multimodal base distributions for continuous-time normalising flows

Shane Josias[*] **and Willie Brink**
Department of Mathematical Sciences
Stellenbosch University
{josias,wbrink}@sun.ac.za

Abstract

We investigate the utility of a multimodal base distribution in continuous-time normalising flows. Multimodality is incorporated through a Gaussian mixture model (GMM) centred at the empirical means of a target distribution's modes. In- and out-of-distribution likelihoods are reported for flows trained with a unimodal and multimodal base distribution. Our results show that the GMM base distribution leads to performance that is comparable to a standard (unimodal) Gaussian distribution for in-distribution likelihoods, but provides the ability to sample from a specific mode in the target distribution, yields generated samples of improved quality, and gives more reliable out-of-distribution likelihoods for low-dimensional input spaces. We conclude that a GMM base distribution is an attractive alternative to the standard base, whose inclusion incurs little to no cost and whose parameterisation may assist with more reliable out-of-distribution likelihoods.

1 Introduction

Normalising flows are a flexible class of generative models that provide exact likelihoods. A discrete-step normalising flow specifies a target distribution $p_x(\boldsymbol{x})$ in terms of an easy-to-sample-from base distribution $p_u(\boldsymbol{u})$, and an invertible transformation $\boldsymbol{u} = T(\boldsymbol{x})$ where $\boldsymbol{u} \sim p_u(\boldsymbol{u})$, by employing the change of variables formula. $T(\boldsymbol{x})$ is defined as a composite function, usually chosen to be a neural network whose architecture is restricted for a tractable log-determinant in the change of variables formula. The continuous-time variant (hereafter referred to as a continuous flow) expresses $\boldsymbol{u} = T(\boldsymbol{x})$ as the solution to an initial value problem (IVP):

$$\frac{d\boldsymbol{z}(t)}{dt} = f_\theta(\boldsymbol{z}_t, t), \qquad t \in [t_0, t_1], \qquad \boldsymbol{z}_0 = \boldsymbol{z}(t_0) = \boldsymbol{x}, \qquad \boldsymbol{z}_1 = \boldsymbol{z}(t_1) = \boldsymbol{u}, \tag{1}$$

and uses a continuous analog of the change of variables formula to determine $\log p_x(\boldsymbol{z}_0)$ (Chen et al., 2018). The function $f_\theta(\boldsymbol{z}_t, t)$ defines a time-dependent vector field describing the transformation dynamics, with learned parameters θ. This formulation circumvents restrictions on the transformation $T(\boldsymbol{x})$ for a tractable log-determinant, at the time-cost of numerically solving the IVP in equation 1. For a chosen base distribution and transformation function, the flow is trained by maximum likelihood. It is sufficient for the base distribution $p_u(\boldsymbol{z}_1)$ to be defined as a standard unimodal Gaussian (Papamakarios et al., 2021; Kobyzev et al., 2020; Dinh et al., 2015, 2017; Kingma and Dhariwal, 2018; Chen et al., 2018; Grathwohl et al., 2019; Finlay et al., 2020; Voleti et al., 2021), however, when trained with the maximum likelihood objective, these models can provide unreliable likelihoods for out-of-distribution data (Nalisnick et al., 2019; Voleti et al., 2021). For instance, a model trained on FashionMNIST may provide high likelihoods for MNIST samples. Moreover, there is no inherent mechanism to model multimodality when $p_u(\boldsymbol{z}_1)$ is chosen to be the standard Gaussian.

[*] Corresponding author.

DLDE-III Workshop in the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Papamakarios et al. (2017); Kirichenko et al. (2020) and Stimper et al. (2022) incorporate multimodality for discrete-step normalising flows by using a Gaussian mixture model (GMM) as base distribution. Improvements are shown for density estimation and semi-supervised image classification. Bearing in mind the cost of numerically solving the IVP in equation 1, we evaluate the prospects of using a GMM base distribution for supervised density estimation in continuous flows. Our work complements existing approaches by reporting both in- and out-of-distribution likelihoods, with a view towards understanding the out-of-distribution failure modes of continuous flows. We further quantitatively show improvements in sample quality. Collectively, our findings warrant further exploration of the benefits and limitations of using multimodal base distributions in continuous flows.

2 Background and methods

Given a class-labelled training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, and letting $\boldsymbol{z}_0 = \boldsymbol{x}_i$, we construct a continuous flow that computes the likelihood $p_x(\boldsymbol{z}_0)$ from

$$\log p_x(\boldsymbol{z}_0) = \log p_u(\boldsymbol{z}_1) + \int_{t_0}^{t_1} \operatorname{Tr}\left(\frac{\partial f_\theta}{\partial \boldsymbol{z}_t}\right) dt. \tag{2}$$

Following Grathwohl et al. (2019), the transformed sample $\boldsymbol{u} = \boldsymbol{z}_1$ and $\log p_x(\boldsymbol{z}_0)$ are solved simultaneously.

GMM base distribution. To incorporate multimodality, we consider the GMM base distribution with a component for each of the K classes in the data:

$$p_u(\boldsymbol{z}_1) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k, I), \tag{3}$$

with $\boldsymbol{\mu}_k$ set to the empirical mean of each class represented in the training set. During training, we evaluate the log-likelihood for each sample $\boldsymbol{x}_i$ by the base component corresponding to class y_i. At test time, the likelihood of a sample is evaluated and weighed across all components as in equation 3. Double precision and the log-sum-exp trick are used to avoid underflow.

Gaussian base distribution. As a baseline, we consider the base distribution $p_u(\boldsymbol{z}_1) \sim \mathcal{N}(\boldsymbol{\mu}, I)$. In our experiments, $\boldsymbol{\mu}$ is set to the empirical mean of the training data, for direct comparison with the GMM case. This deviates slightly from existing literature where the base distribution is usually defined to be the zero-mean isotropic Gaussian. Different parameterisations (uni- or multimodal) may impact performance, but further investigation into this is left for future work.

These two base distributions will first be evaluated in a 2-dimensional setting, and then on image datasets of increasing complexity.

Cascading moons dataset. Figure 1 shows a 2D cascading moons dataset with 6 classes. We construct two training sets, $\mathcal{D}_{\{0,1\}}$ and $\mathcal{D}_{\{4,5\}}$, where the subscripts represent the set of in-distribution modes the model is trained on. In both cases, the remaining classes serve as out-of-distribution data for testing. The dynamics function $f_\theta(\boldsymbol{z}_t, t)$ is implemented as a planar normalising flow (Rezende and Mohamed, 2015) where a hypernetwork (Ha et al., 2017) is used for time-dependency in the planar flow's parameters.

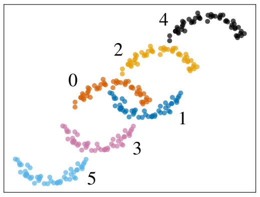

Figure 1: The 2D cascading moons dataset.

Image datasets. We consider three image datasets: (1) InstanceMNIST, where the training set contains single samples from class 0 and from class 1, copied with uniform noise of up to 20% of the maximum pixel value; (2) SubsetMNIST, where the training set has 1280 different samples from each of the classes 0 and 1; and (3) FashionMNIST, where the training set consists of 1280 samples from each of the 10 classes. For computational convenience, all images are resized from 28 × 28 to 16 × 16. The-out-of-distribution data for models trained on InstanceMNIST and SubsetMNIST consists of samples from classes 2 to 9, constructed in the same way as each respective training set, and an additional set of linear interpolations of samples from the two training classes. The

MNIST test set serves as out-of-distribution data for models trained on FashionMNIST, to test whether continuous flows with a multimodal base distribution still erroneously identifies MNIST as in-distribution data (Nalisnick et al., 2019; Voleti et al., 2021). The dynamics function $f_\theta(\boldsymbol{z}_t, t)$ is implemented as a fully convolutional neural network with time concatenated to the input of each layer, similar to Grathwohl et al. (2019). The Jacobian trace in equation 2 is calculated using the Hutchinson trace approximation (Grathwohl et al., 2019; Finlay et al., 2020), and weight normalisation is included when training on FashionMNIST.

We report on bits per dimension (bits/dim) scores for the in- and out-of-distribution data. A higher bits/dim score implies lower average likelihood. The Fréchet inception distance (FID) (Heusel et al., 2017) is also reported as a quantitative measure of generated sample quality.

3 Results

Figure 2 shows comparable performance in modelling in-distribution likelihoods between the standard Gaussian and GMM base distributions, for the cascading moons dataset, indicating that there is no degradation in in-distribution likelihood performance in this low-dimensional setting. The results also indicate that the GMM base improves over the standard base by correctly assigning higher bits/dim to out-of-distribution data. Moreover, the bits/dim scores are correlated with how far the test distribution modes are from the training distribution. It may be worth analysing the learned dynamics for the GMM base to determine what characteristics of its parameterisation allows for these improved out-of-distribution scores.

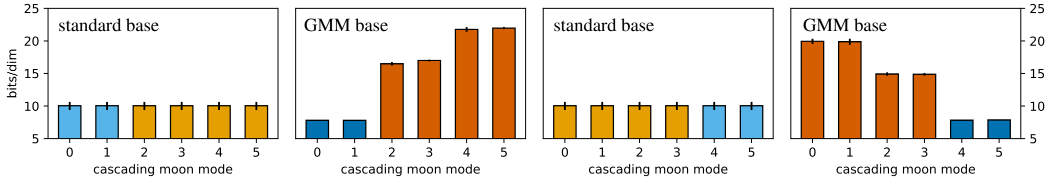

Figure 2: Bits/dim scores per class in the cascading moons dataset. Blue indicates training set modes (in-distribution) and orange indicates out-of-distribution modes. Black vertical lines represent standard deviations over a few runs.

Nalisnick et al. (2019) suggest that flows fail at modelling out-of-distribution likelihood when the test distribution is contained within the training distribution, i.e. when the two sets have similar means but the test set has smaller variance. We also observe this behaviour in the low-dimensional setting, when training on modes 4 and 5 with a standard base. The GMM base, however, performs desirably. The hypothesis that data from all modes collapse to the unimodal base distribution in the solution to the IVP in equation 1 serves as motivation for using a multimodal base distribution.

Both variants of the base distributions lead to similar performance on the InstanceMNIST data, as seen in Figure 3. Again, there is no performance degradation for in-distribution likelihoods from the GMM base. In these experiments, interpolated samples between the training modes are identified as in-distribution. We expect as much from the standard base distribution, given that it is centred at the

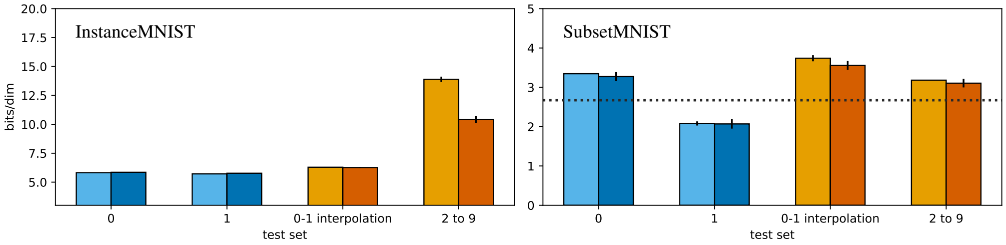

Figure 3: Bits/dim scores for in- and out-of-distribution sets, for the InstanceMNIST (left) and SubsetMNIST (right) experiments. Models are trained on samples from classes 0 and 1, and tested on interpolated samples as well as samples from classes 2 to 9. Lighter and darker shades correspond to the standard and GMM bases, respectively.

interpolated modes (similar to the cascading moons experiment). But the same behaviour occurs for the GMM base distribution. It is possible that the dynamics function moves data from the interpolated mode to regions of high probability for either of the modes, suggesting that parameterisation of the base distribution modes might be an important consideration. Both models correctly identify data from classes 2 to 9 as out-of-distribution. The InstanceMNIST set contains data in a relatively small hypercube centred at a single sample from class 0 and a single sample from class 1, leading to models that fit the training set very well.

Results in Figure 3 for SubsetMNIST again show that continuous flows trained with a GMM base can model in-distribution data adequately. More interestingly, it appears that both models (standard base and GMM base) assign similar likelihoods to out-of-distribution data as they do for in-distribution data from class 0, suggesting that harder-to-model in-distribution classes may affect mean bits/dim scores typically reported in the literature. When considering the mean bits/dim (dotted) line, it is not clear whether the data from class 0 should be regarded as out-of-distribution, or whether the out-of-distribution data from classes 2 to 9 should be regarded as in-distribution.

Table 1: Mean bits/dim for models trained on FashionMNIST.

	Standard	GMM
FashionMNIST	3.87	3.86
MNIST	3.21	3.22

Models trained on FashionMNIST are tested on FashionMNIST test samples (in-distribution) and MNIST test samples (out-of-distribution). In-distribution likelihoods are similar for models trained with the standard and GMM bases, as indicated by the bits/dim scores in Table 1 (standard deviations over a few runs are in the order of 10^{-2}). Both models assign higher likelihoods to out-of-distribution data, similar to results from the literature (Nalisnick et al., 2019; Voleti et al., 2021). A multimodal base distribution does not seem to assist with out-of-distribution detection, and more investigation is needed. Training on feature representations may circumvent the out-of-distribution failure (Kirichenko et al., 2020). It could be that linear separability is important for flows to enable out-of-distribution detection. Perhaps there may be an appropriate parameterisation of the GMM so that flow trajectories behave well for in-distribution data, while pushing out-of-distribution data to regions of low probability.

Table 2: FID for generated images from models trained on the indicated sets.

	Standard	GMM
InstanceMNIST	97.59	84.68
SubsetMNIST	144.26	137.20
FashionMNIST	80.73	77.25

Table 2 provides quantitative evidence that a GMM base distribution can lead to better generated samples, as measured by FID. This echoes what has been shown qualitatively for discrete-step normalising flows (Dinh et al., 2017; Stimper et al., 2022). We acknowledge that sample quality and likelihood performance are largely independent (Theis et al., 2016), and so this result demonstrates the benefit of using a multimodal base distribution only for applications where sample quality is important.

Importantly, in the case of class-labelled training data, we observe that using a GMM base distribution does not increase the time-cost of solving the IVP in equation 1. Comparable likelihoods for in- and out-of-distribution data, a comparable time-cost and improved sample quality provide support for the use of a multimodal base distribution in continuous-time normalising flows.

4 Conclusion and future work

We investigated the utility of a GMM base distribution for continuous flows, where the data transformation is defined as the solution to an initial value problem. We showed that continuous flows trained with a GMM base distribution can generate better quality samples at no additional cost to the training or inference process. Comparable in-distribution likelihoods, and reliable out-of-distribution likelihoods for low-dimensional input spaces, further motivate for a multimodal base distribution as a simple alternative to the standard Gaussian.

Further analysis is warranted into the relationship between the parameterisation of the base distribution and the learned dynamics for reliable out-of-distribution likelihoods. Indeed, labelled data simplifies the parameterisation of the GMM, but it is still an open question as to whether unlabelled data can be handled in a sensible way. Learning the parameters of the GMM is another avenue to investigate.

Acknowledgement

This work is based on research supported by the National Research Foundation of South Africa (grant number 138341).

References

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in Neural Information Processing Systems*.

Dinh, L., Krueger, D., and Bengio, Y. (2015). NICE: non-linear independent components estimation. *International Conference on Learning Representations, Workshop Track Proceedings*.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using real NVP. *International Conference on Learning Representations, Conference Track Proceedings*.

Finlay, C., Jacobsen, J., Nurbekyan, L., and Oberman, A. M. (2020). How to train your neural ODE: the world of Jacobian and kinetic regularization. *International Conference on Machine Learning*.

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., and Duvenaud, D. (2019). FFJORD: free-form continuous dynamics for scalable reversible generative models. *International Conference on Learning Representations, Conference Track Proceedings*.

Ha, D., Dai, A. M., and Le, Q. V. (2017). HyperNetworks. *International Conference on Learning Representations, Conference Track Proceedings*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*.

Kingma, D. P. and Dhariwal, P. (2018). Glow: generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems*.

Kirichenko, P., Izmailov, P., and Wilson, A. G. (2020). Why normalizing flows fail to detect out-of-distribution data. *Advances in Neural Information Processing Systems*.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2020). Normalizing flows: an introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979.

Nalisnick, E. T., Matsukawa, A., Teh, Y. W., Görür, D., and Lakshminarayanan, B. (2019). Do deep generative models know what they don't know? *International Conference on Learning Representations, Conference Track Proceedings*.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680.

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. *Advances in Neural Information Processing Systems*.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. *International Conference on Machine Learning*.

Stimper, V., Schölkopf, B., and Hernández-Lobato, J. M. (2022). Resampling base distributions of normalizing flows. *International Conference on Artificial Intelligence and Statistics*.

Theis, L., van den Oord, A., and Bethge, M. (2016). A note on the evaluation of generative models. *International Conference on Learning Representations, Conference Track Proceedings*.

Voleti, V., Finlay, C., Oberman, A., and Pal, C. (2021). Multi-resolution continuous normalizing flows. *arXiv preprint arXiv:2106.08462*.