Networks for Variations: A Review of Normalizing Flows for Bayesian Variational Inference

Søren Winkel Holm

June 23, 2022

Introduction

In many technical fields, modelling complex systems is in recent years achieved using Deep Learning (DL) instead of setting up domain-suitable inferential statistical models [BAK18; Bre01]. The success of DL can be attributed to the potential for using similar algorithms to achieve high prediction accuracy across different big data problems [Par15]. However, a need for moving these high-accuracy, black box methods towards more robustness and explainability has been highlighted. Seeking this goal, methods have been developed for characterising complete distributions of model predictions, parametrizations or training data instead of only focusing specific realisations of these. This distributional view of DL is used in Deep Generative Modelling (DGM) and in the wider context Bayesian Machine Learning (BML). For both the task of DGM specifically and the general task of approximating posterior distributions in BML, robust and general methods for constructing complex distributions are needed. Normalizing Flows (NF's) specify a scalable mechanism allowing for the representation of arbitrary distributions. This method is here reviewed with a focus on its' relevance for DL.

Fundamental Concepts

Let

$$\mathbf{Z} \in \mathbb{R}^2 \sim p_{\mathbf{Z}}, \tag{1}$$

where $p_{\mathbf{Z}}$ is known and analytically tractable distribution, e.g. $p_{\mathbf{Z}} = \mathcal{N}$. Now using a composition of N bijective functions $\mathbf{g} = \mathbf{g}_N \circ \cdots \circ \mathbf{g}_1$ with inverse $\mathbf{f} = \mathbf{f}_N \circ \cdots \circ \mathbf{f}_1$ and Jacobian $\mathbf{Dg}$, set $\mathbf{W} = \mathbf{g}(\mathbf{Z})$ giving the density of $\mathbf{W}$

$$\mathbf{g}_\star p_{\mathbf{Z}}(\mathbf{w}) = p_{\mathbf{W}}(\mathbf{w}) = \frac{p_{\mathbf{Z}}\left(\mathbf{f}(\mathbf{w})\right)}{\left|\det \mathbf{Dg}\left(\mathbf{f}(\mathbf{w})\right)\right|}. \tag{2}$$

In the context of NF's, $\mathbf{g}_\star p_{\mathbf{Z}}$ is named the *pushforward* of the base density $p_{\mathbf{Z}}$. $\mathbf{g}_\star p_{\mathbf{Z}}$ pushes the simple density $p_{\mathbf{Z}}$ to a possibly arbitrarily complex distribution which

is called flow in the *generative direction* [KPB21] as

$$\mathbf{z} \sim p_{\mathbf{Z}} \wedge \mathbf{w} = g(\mathbf{z}) \Rightarrow \mathbf{w} \sim \mathbf{g}_\star p_{\mathbf{Z}}. \quad (3)$$

Inversely, $\mathbf{f}$ moves density towards the simple distribution, a process called flow in the *normalizing direction* [KPB21].

Using this construction, arbitrarily complex distributions $p_{\mathbf{W}}$ can provenly be represented [BKM07], but the functions are only considered NF's if $\mathbf{g}_i$, $\mathbf{f}_i$ and the Jacobian determinant are easy to compute [KPB21] e.g. using

$$|\det \mathbf{Dg}\left(\mathbf{f}(\mathbf{w})\right)|^{-1} = \quad (4)$$
$$\left|\prod_i^N \det \mathbf{Df}_i\left(\mathbf{f}_{i+1} \circ \cdots \circ \mathbf{f}_N(\mathbf{w})\right)\right|.$$

$\mathbf{g}$ may have a parametrization ϕ, resulting in the pushforward being parameter dependant $\mathbf{g}_\star p_{\mathbf{Z}}(\mathbf{w}|\phi)$.

Using (2), NF's allows for density evaluation and using (3) for sampling. The first quality makes the method relevant for Variational Inference (VI) used in BML for approximating $p = p(\mathbf{w}|\mathcal{D})$ as

$$q^\star = \text{argmin}_{q \in \mathcal{Q}} \mathbb{D}_{KL}[q||p] \quad (5)$$

where $\mathbb{D}_{KL}$ is the Kullback-Leibler divergence (KL) and $\mathcal{Q}$ is the variational family of possible approximations [BKM16]. Minimization of KL corresponds to maximization of the evidence lower bound (ELBO) which is, assuming this family is parametrized with ϕ,

$$\mathcal{L}(\phi) = \mathbb{E}_{q|\phi}[\ln p(\mathbf{w}, \mathcal{D})] - \mathbb{E}_{q|\phi}[\ln q(\mathbf{w}|\phi)]. \quad (6)$$

To optimize this without model-dependant derivations, gradient ascent on the $\mathcal{L}$ is carried out resulting in Black-box Variational Inference (BBVI). Here, computing gradients of the form $\nabla_\phi \mathbb{E}_{q|\phi}[h(\mathbf{w})]$ is required. If NF's are used such that $q(\mathbf{w}|\phi) = \mathbf{g}_\star p_{\mathbf{Z}}(\mathbf{w}|\phi)$, the gradients can be computed using the reparametrization trick [RM15]

$$\nabla_\phi \mathbb{E}_{q|\phi}[h(\mathbf{w})] = \nabla_\phi \mathbb{E}_{p_{\mathbf{Z}}}[h(\mathbf{g}(\mathbf{z}|\phi))]. \quad (7)$$

An alternative use for NF's is directly modelling data as a type of density estimation. Here, data likelihood is

$$\ln p(\mathcal{D}|\phi) = \sum_{i=1}^M \ln \mathbf{g}_\star p_{\mathbf{Z}}(\mathbf{y}_i|\phi) =$$
$$\sum_{i=1}^M \left(\ln p_{\mathbf{Z}}(\mathbf{f}(\mathbf{y}_i|\phi)) + \ln |\det \mathbf{Df}(\mathbf{y}_i|\phi)|\,,\right).$$

This model is generative using (3) and can be fitted using Maximum Likelihood Estimation (MLE).

State of the Art

- Introduction recently, compare with standard bbvi

- Examples of nfs

Open Problems

- Choice of base density?
- What flows are efficient?
- Discrete distributions?

References

[BAK18] Danilo Bzdok, Naomi S. Altman, and Martin Krzywinski. "Points of Significance: Statistics versus machine learning". In: *Nature Methods* 15 (2018), pp. 233–234.

[BKM07] Vladimir Bogachev, Alexander Kolesnikov, and Kirill Medvedev. "Triangular transformations of measures". In: *Sbornik: Mathematics* 196 (Oct. 2007), p. 309. DOI: 10.1070/SM2005v196n03ABEH000882.

[BKM16] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112 (2016), pp. 859–877.

[Bre01] Leo Breiman. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". In: *Statistical Science* 16.3 (2001), pp. 199–231. DOI: 10.1214/ss/1009213726. URL: https://doi.org/10.1214/ss/1009213726.

[KPB21] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. "Normalizing Flows: An Introduction and Review of Current Methods". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (2021), pp. 3964–3979. DOI: 10.1109/TPAMI.2020.2992934.

[Par15] Roger Parloff. "Why Deep Learning Is Suddenly Changing Your Life". In: *Fortune* (Sept. 28, 2015). URL: https://fortune.com/longform/ai-artificial-intelligence-deep-machine-learning/ (visited on 06/10/2022).

[RM15] Danilo Jimenez Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: *ICML*. 2015.