# Advanced Topics in Machine Learning
# Group 12 Project Report

Reproducing results from "Variational Inference with Normalizing Flows" [Danilo Jimenez, Rezende Shakir Mohamed]


1032626, 1034125, 1034129, 1036969

University of Oxford

# Introduction

The idea of generative models is one of the most promising approaches to Machine Learning tasks. Many such models have been used in recent years to solve the problem of inference of a posterior distribution, given some expert knowledge and assumptions on the prior distribution and using observed data to adapt these. The main obstacle in attempting to generate the true posterior is its computational intractability. Consequently, approximation methods have been employed, aiming to make computation more efficient while at the same time achieving a close estimate of the true posterior.

The paper focusses on the method of variational inference. Essentially, this converts the inference problem to an optimization one where, given a family of posterior approximations, we look for the distribution that is the closest to the true one. To measure this proximity between approximations and the true posterior, we use the Kullback-Leibler divergence. There has been lots of research in the classes of such, efficiently calculatable, approximation families, such as mean-field approximations, but all of these prove to be of limited expressiveness, leading to solutions that are far from the true posterior. In fact, there is strong evidence showing that more expressive classes of posteriors would result in better performance of variational inference, which is what motivates the work of the paper.

To address the above issue, the authors propose constructing the approximations through a normalizing flow, which applies a sequence of invertible transformations to a simple initial density until it achieves a desired level of complexity. The main types of flows considered and used in the experiments are:

1. Invertible, linear-time flows, such as planar and radial flows. 2. Finite, volume-preserving flows, such as the Non-linear Independent Components Estimation (NICE).

The goal of both of the above is to provide a transformation to the original density that ends up being rich and faithful, while at the same time allowing for efficient computation of the determinant. The first kind of flow allows for computing the logdet-Jacobian in linear (in the number of dimensions) time, while the second one gives a Jacobian with a zero upper triangular part and therefore a determinant of 1. Thus, both can be efficiently implemented and used together with a Deep Latent Gaussian Model and an inference network to solve our problem.

In addition to the practical implementation and the experiments, the paper gives proof that infinitesimal flows, which are normalizing flows whose length tends to infinity, can define a class of posteriors that is asymptotically able to recover the true posterior distribution, solving therefore one of the most important problems of variational inference. This is another good indicator that the method should perform well in the finite case.

- TODO: describe in more depth, after we have results The paper's experiments on the binarized MNIST dataset aim to demonstrate how variational inference with normalizing flows can outperform previous methods, compare the performance of different types of flows, and discuss the relation between the length of the flow and the quality of the resulting approximation.

- TODO: after we finish the extension In our report, we replicate the original contributions of the paper by running the same experiments on MNIST. Additionally, we consider another type of flow, the invertible convolutional flow .... Finally, we provide an open-source implementation of the above in our Github repository: `https://github.com/ATML-Group-12/normalising_flows`

# Normalizing Flows

In the attempt to generate more complex families of posterior distributions, which should allow for better estimates of the true posterior, the paper introduces the idea of normalizing flows. A normalizing flow is a sequence of invertible transformations applied to a simple initial distribution (for example a unit gaussian), resulting in a complex distribution, whose density we can efficiently evaluate by inverting the flow and keeping track of the Jacobian of the composition of transformations, using the change of variable theorem.

## Definition

More formally, consider a member of the flow's sequence of transformations to be an invertible, smooth mapping $f : \mathbb{R}^d \to \mathbb{R}^d$. Let $\mathbf{y} = f(\mathbf{z})$ be the outcome of applying the transformation to a random variable $\mathbf{z}$ with distribution $q(\mathbf{z})$. Then, using the change of variable theorem, the resulting density is:

$$q(\mathbf{y}) = q(\mathbf{z}) \left| \det \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right| = q(\mathbf{z}) \left| \det \frac{\partial f(\mathbf{z})}{\partial \mathbf{z}} \right|^{-1}$$

where the second equality follows from the inverse function theorem and the definition of $\mathbf{y}$.

We can then proceed to build more complicated densities by systematically composing multiple transformations such as the one above, since the composition of invertible transformations remains an invertible transformation. To simplify notation, we define the log-density $\ln q_K$ resulting from applying a sequence of $K$ transformations $f_1, f_2, ..., f_K$ to a random variable $\mathbf{z_0}$ with an initial distribution $q_0$ as:

$$\mathbf{z_K} = f_K \circ ... \circ f_2 \circ f_1(\mathbf{z_0})$$

$$\ln q_K(\mathbf{z_K}) = \ln q_0(\mathbf{z_0}) - \sum_{k=1}^{K} \ln \left| \det \frac{\partial f_k}{\partial \mathbf{z_{k-1}}} \right|$$

where the first equation represents the sequence of random variables generated by the flow. The normalizing flow is then defined as the path of the successive distributions $q_k$.

## Law of the Unconscious Statistician

A useful property of these transformations is what is known as the law of the unconscious statistician (LOTUS). The LOTUS refers to the fact that we can evaluate expectations with respect to the transformed density $q_K$ without explicitly knowing it, by expressing any such expectation as:

$$\mathbb{E}_{q_K} = \mathbb{E}_{q_0}[h(f_K \circ ... \circ f_2 \circ f_1(\mathbf{z_0}))]$$

This is an expectation over the known density $q_0$, which does not require computation of the logdet-Jacobian terms when $h(\mathbf{z})$ does not depend on $q_K$.

## Invertible Linear-time Transformations

Notice that a naive choice of these transformations would lead to a $O(d^3)$ complexity to compute the determinant of the Jacobian. Therefore, the classes of flows studied in our report are all based on the idea of having an efficient way to calculate this determinant. We begin by investigating two types of linear flows, namely planar and radial flows.

### Planar Flows

Consider the following class of transformations:

$$f(\mathbf{z}) = \mathbf{z} + \mathbf{u}h(\mathbf{w}^T\mathbf{z} + b)$$

where $\lambda = \{\mathbf{w} \in \mathbb{R}^D, \mathbf{u} \in \mathbb{R}^D, b \in \mathbb{R}\}$ are free parameters and $h(\cdot)$ is a smooth, element-wise non-linearity. We can use the matrix determinant lemma to calculate the logdet-Jacobian term in linear time, yielding:

$$\left|\det \frac{\partial f}{\partial \mathbf{z}}\right| = \left|1 + \mathbf{u}^T\psi(\mathbf{z})\right|$$

where $\psi(\mathbf{z}) = h'(\mathbf{w}^T\mathbf{z} + b)\mathbf{w}$ and $h'$ is the derivative of the function $h$. Using this, we can now substitute in the formula for $\ln q_K$ to get the following closed form for planar flows:

$$\ln q_K(\mathbf{z_K}) = \ln q_0(\mathbf{z}) - \sum_{k=1}^{K} \ln \left|1 + \mathbf{u_k}^T\psi_k(\mathbf{z_{k-1}})\right|$$

The name planar comes from the fact that the above transformation applies a series of contractions and expansions in the direction perpendicular to the $\mathbf{w}^T\mathbf{z} + b = 0$ hyperplane to the initial density $q_0$.

### Radial Flows

The other family of invertible linear-time transformations studied is defined by the following equation:

$$f(\mathbf{z}) = \mathbf{z} + \beta h(\alpha, r)(\mathbf{z} - \mathbf{z_0})$$

where $r = |\mathbf{z} - \mathbf{z_0}|$, $h(\alpha, r) = 1/(\alpha + r)$, and the set of parameters is $\lambda = \{\mathbf{z_0} \in \mathbb{R}^D, \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}\}$. The time-complexity of the computation of the determinant of this class is also linear:

$$\left|\det \frac{\partial f}{\partial \mathbf{z}}\right| = [1 + \beta h(\alpha, r)]^{d-1}[1 + \beta h(\alpha, r) + \beta h'(\alpha, r)]$$

This transformation results in radial contractions and expansions around the reference point $\mathbf{z_0}$, hence the name radial flows.

# NICE

Using the idea of flows, the authors view the coupling layers in the NICE paper as a flow. A coupling layer $f$ is a neural network layer with easy to compute inverse and a trivial Jacobian. For an input vector $\mathbf{z} \in \mathbb{R}^D$, we have

$$f(\mathbf{z}) = (\mathbf{z}_A, g(\mathbf{z}_B, h(\mathbf{z}_A)))$$
$$f^{-1}(\mathbf{z}) = (\mathbf{z}_A, g^{-1}(\mathbf{z}_B, h(\mathbf{z}_A)))$$

where $(A, B)$ is a partition of $\{1, 2, \ldots, D\}$, $h$ is an arbitrary function with input size $|A|$, and $g$ is a coupling law, a function that is invertible for the first argument given the second. In the paper, $h$ is a neural network and $g(a, b) = a + b$, so the Jacobian is just the identity matrix. Since the determinant of the Jacobian is 1, the authors of (variation paper) classify it as a volume-preserving flow. The authors of the NICE paper do point out this issue, and introduce a diagonal scaling layer as the last layer of their models. This approach was not used in the original paper. Moreover, the authors introduced variants NICE-ortho and NICE-perm instead of the original directly, where random orthogonal and permutation matrices are applied to $\mathbf{z}$ before the partition.

We note that the authors of the original paper have not made the choice of $h$ clear in the experiments. For the replicating experiment 6.1, we have chosen $h$ to be a ReLU network of 4 layers with hidden dimensions of 2.

# Conclusion

TODO: Conclusion
- critical evaluation of our work
- what we could've done with more time/resources

# References


[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[2] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation, 2014.

[3] Matt Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference, 2012.

[4] Mahdi Karami, Dale Schuurmans, Jascha Sohl-Dickstein, Laurent Dinh, and Daniel Duckworth. Invertible convolutional flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[5] Yann LeCun and Corinna Cortes. The MNIST database of handwritten digits. 1998.

[6] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2015.