
Variational Inference with Non-invertible Flow

Anonymous Author(s)

Affiliation

Address

email

Abstract

We explore the possibility of extending normalizing flow with non-invertible transformations for variational inference. We show that the probability density function of the resulting distribution can be derived analytically, but it depends on all the real roots of the transformation equation. To address this limitation, we develop non-invertible flow (NFW), a variational inference method with non-invertible transformations by approximating the degenerate conditional distribution of transformation with a normal distribution with a small variance. This brings a close connection to semi-implicit variational inference (SIVI) methods. Unlike SIVI methods, our variational lower bound does not require a large sample size and does not need to run a Markov Chain Monte Carlo method during variational inference. We demonstrate NFW on several tasks, including variational autoencoder, image imputation and Bayesian neural network regression, and show that it can capture complex distributions and deliver good performance on various tasks.

1 Introduction

Variational inference (VI) is an important family of methods for scaling probabilistic modeling for real world problems. VI converts an intractable posterior estimation of Bayesian inference into an optimization problem, in which the task is to search for the parameters of a chosen distribution that minimize the Kullback-Leibler (KL) divergence between the chosen parametric distribution and the true posterior distribution. The objective function of optimization is often formulated as optimizing a variational lower bound of the log marginal likelihood, which is also called evidence lower bound (ELBO). Two major challenges of VI are (1) calculation of the integral in ELBO and (2) expressiveness of variational distributions. A popular approach is to address the problem of the integral calculation is to approximate the ELBO with Monte Carlo samples. By using reparameterizable distributions, a lower variance estimation is achievable with a relatively small number of samples. This requires a rich family of reparameterizable variational distributions. Many works focus on developing reparameterizable representations of known distributions [Ruiz et al., 2016, Naesseth et al., 2016, Figurnov et al., 2018]. This significantly extends the choices of variational distributions, however, for large scale probabilistic models, a good variational distribution often needs to be high dimensional, highly correlated among dimensions, and does not make with known distribution families.

Normalizing flow (NF), proposed by Jimenez Rezende and Mohamed [2015], constructs complex probabilistic distributions via transforming random variables from a simple probability distribution with a known probability density function (PDF). By restricting the transformation to be invertible, a resulting distribution can be used for variational inference, because (1) it is reparameterizable, as a sample can be formulated as a function of the parameters of transformation, (2) the PDF can be calculated in closed form. The main limitation of NF is the computational expensive PDF calculation, in which the log determinant of the Jacobian of the transformation needs to be computed. This is impractical for the distributions with thousands of dimensions, e.g., the posterior of a Bayesian

neural networks (BNN). Many follow-up works [Kingma et al., 2016, Dinh et al., 2017, Huang et al., 2018] address this problem by carefully constructing the transformation, of which the Jacobian is a triangular matrix. This allows efficient computation of the log determinant of Jacobian, but is at the cost of enforcing an order of dependency among dimensions and being slow at generating samples, which is because the samples have to go through an autoregressive model.

In this paper, we explore the possibility of extending NF with non-invertible transformations. The major challenge of using non-invertible transformations is to calculate the PDF of the resulting distribution. The change of variable formula used to derive the PDF of NF is no longer applicable. We show that, if the dimensionality of the initial distribution is greater than or equal to the dimensionality of the resulting distribution, the PDF can be derived analytically, but is not practical because it depends on all the real roots of the transformation equation. To address this limitation, we approximate the conditional distribution of the transformation, which is a Dirac delta distribution, with a normal distribution with a small variance. With the approximated conditional distribution, we develop non-invertible flow (NFW), a variational inference method using a non-invertible transformation, in which a variational lower bound is derived by introducing an auxiliary distribution. The formulation of connecting a simple random variable to the target variable with a non-invertible function in variational posterior is closely related to semi-implicit variational inference (SIVI) methods [Yin and Zhou, 2018, Titsias and Ruiz, 2019, Molchanov et al., 2019]. Unlike the SIVI methods, our variational lower bound works well with a small number of samples and does not need to run a Markov Chain Monte Carlo method during variational inference. We show that the auxiliary distribution in NFW produces a similar regularization effect as the log determinant in NF, which demonstrates a strong connection between the NF-based methods and the SIVI methods.

2 Variational inference with implicit distribution

Consider a set of observed variables \mathbf{y} and a set of latent variables \mathbf{x} , of which a probabilistic model is defined in terms of a conditional distribution over the observed variables $p(\mathbf{y}|\mathbf{x})$ and a prior distribution over the latent ones $p(\mathbf{x})$. VI is often applied to estimate the intractable posterior distribution $p(\mathbf{x}|\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})/p(\mathbf{y})d\mathbf{x}$. VI solves this intractable posterior estimation by converting the integral calculation into an optimization problem, where we define a parametric distribution as the variational posterior and optimize the parameters of the variational posterior to match the true posterior distribution. This optimization is often formulated as minimizing the KL divergence between the variational posterior and the true posterior,

$$\text{KL}(q_\theta(\mathbf{x}) \| p(\mathbf{x}|\mathbf{y})) = \langle \log p(\mathbf{y}, \mathbf{x}) - \log q_\theta(\mathbf{x}) \rangle_{q_\theta(\mathbf{x})} - \log p(\mathbf{y}), \quad (1)$$

where $q_\theta(\mathbf{x})$ is the variational posterior and θ are the parameters of the variational posterior. Note that the marginal likelihood term $\log p(\mathbf{y})$ is independent of $q_\theta(\mathbf{x})$. VI focuses on minimizing the first term $\langle \log p(\mathbf{y}, \mathbf{x}) - \log q_\theta(\mathbf{x}) \rangle_{q_\theta(\mathbf{x})}$, which is often referred to as variational lower bound or ELBO.

Despite the fact that the form of the variational posterior is chosen by hand, for most of sophisticated probabilistic models, it is hard to find a variational posterior distribution that gives a closed-form variational lower bound. Instead of simplifying the model and variational posterior for a closed form variational lower bound, stochastic variational inference (SVI) optimizes the variational posterior with respect to a Monte Carlo estimation of the variational lower bound,

$$\mathcal{L} \approx \frac{1}{K} \sum_{i=1}^K \log p(\mathbf{y}, \mathbf{x}_i) - \log q_\theta(\mathbf{x}_i), \quad \mathbf{x}_i \sim q_\theta(\mathbf{x}). \quad (2)$$

This allows us to use more sophisticated probabilistic models and variational posteriors. The variational posterior is typically optimized via gradient optimization. To directly estimate the gradient of a variational posterior from (2), the variational posterior needs to (1) have a closed form PDF, (2) be reparameterizable. Note that in (2), to estimate the gradient of the variational parameters with respect to the variational lower bound, the gradient not only needs to be estimated from the log PDF $\log q(\mathbf{x})$ but also needs to chain through the individual samples, *i.e.*,

$$\frac{\partial \mathcal{L}}{\partial \theta} = -\frac{\partial \log q(\mathbf{x}_i)}{\partial \theta} + \sum_{i=1}^M \frac{\partial \mathcal{L}}{\partial \mathbf{x}_i} \frac{\partial \mathbf{x}_i}{\partial \theta}. \quad (3)$$

84 This requires the random variable to be defined as a transform of a standard probability density. For
 85 example, a normally distributed variable can be written as $x = \mu + \sigma\epsilon$, where μ and σ^2 are the mean
 86 and variance of the normal distribution and ϵ follows a normal distribution with zero mean and unit
 87 variance $\epsilon \sim \mathcal{N}(0, 1)$. This is known as the explicit reparameterization [Kingma and Welling, 2014].

88 2.1 Change of Variable

89 The reparameterization trick allows us to build low variance estimators for a subset of known
 90 probabilistic distribution for SVI, among which the scalar normal distribution remains the most
 91 practical choice. This enables wide usage of mean field approximation in SVI, $q(\mathbf{x}) = \prod_{k=1}^N q(x_k)$.
 92 However, it is known that mean field is a poor assumption for the random variables that are strongly
 93 correlated in the posterior.

Normalizing flow, introduced by Jimenez Rezende and Mohamed [2015], provides a powerful
 approach to generate an complex probability distribution by applying the change of variable formula.
 It starts with an initial random variable $\mathbf{z}_0 \in \mathbb{R}^D$ and applies the transformation functions f_1, \dots, f_T ,

$$\mathbf{z}_0 \sim q(\mathbf{z}_0), \quad \mathbf{z}_t = f_t(\mathbf{z}_{t-1}), \quad \forall t = 1, \dots, T$$

94 where $g(\mathbf{z}_0)$ is the PDF of $q(\mathbf{z}_0)$. Drawing a sample from the resulting distribution $q(\mathbf{z}_T)$ is straight-
 95 forward, which can be done by drawing a sample from the initial distribution and applies the series
 96 of transformation functions. All the transformation functions need to be invertible. According to the
 97 change of variable formula, it is known that the probability density function of the target variable can
 98 be formulated as

$$q(\mathbf{z}_T) = g(f_1^{-1} \circ \dots \circ f_T^{-1}(\mathbf{z}_T)) \prod_{t=1}^T \left| \frac{\partial f_t^{-1}(\mathbf{z}_t)}{\partial \mathbf{z}_t} \right|. \quad (4)$$

99 Sampling from the resulting distribution and the evaluation of the original probability density function
 100 can both be done efficiently. However the computation of the Jacobian of the inverse transformation
 101 can be expensive when the dimensionality of the probability distribution gets high. Kingma et al.
 102 [2016] propose inverse autoregressive flow (IAF) to tackle the computational challenge of the above
 103 determinant by constructing the transformation function of which the Jacobian is an triangular matrix.
 104 This reduces the complexity of the calculation from $O(D^2)$ to $O(D)$. Such a transformation is
 105 restrictive, as only the latter transformation can depend on previous dimensions but not the other way
 106 around. The computation suffers from the common limitation of autoregressive network: slow to
 107 evaluation and need additional tricks for taking care of the gradients.

108 2.2 Non-invertible Transformation

109 An obvious extension to NF is to consider non-invertible transformations. However, the challenge is
 110 that the PDF formula (4) is no longer applicable. *What is the PDF of the target random variable after
 111 a non-invertible transformation?*

112 We start off with looking into how the PDF is derived for 1D cases. Following the rule of marginal-
 113 ization, the probability density function of the target variable x that is transformed from an initial
 114 random variable z with the function f is defined as

$$q(x) = \int q(x|z)q(z)dz = \int \delta(x - f(z))g(z)dz = g(f^{-1}(x)) \left| \frac{\partial f^{-1}(x)}{\partial x} \right|, \quad (5)$$

115 where the PDF of z is $g(z)$ and x has a deterministic relation with z : $x = f(z)$. In this case, the
 116 conditional distribution $q(x|z)$ is degenerate, which can be represented by a Dirac delta function. The
 117 resulting degenerate PDF returns an infinite density at the location $f(z)$ and zero everywhere else,
 118 *i.e.*, $q(x|z) = \delta(x - f(z))$. If f is invertible, using the properties of Dirac delta function [Arfken,
 119 1985], the PDF $q(x)$ can be derived in closed form as shown in (5), which leads to (4).

120 If f is non-invertible, the PDF can also be derived in closed form,

$$q(x) = \int \sum_{i, f(z_i)=x, f'(z_i) \neq 0} \frac{\delta(z - z_i)}{|f'(z_i)|} g(z) dz = \sum_{i, f(z_i)=x, f'(z_i) \neq 0} \frac{g(z_i)}{|f'(z_i)|}, \quad (6)$$

121 where $\{z_i\}$ are all the real roots of the equation $x = f(z)$. Fig. 1a shows an example of transforming
 122 a standard normal distribution $z \sim \mathcal{N}(0, 1)$ with a quadratic function $x = z^2 + 4$. Following (6), the

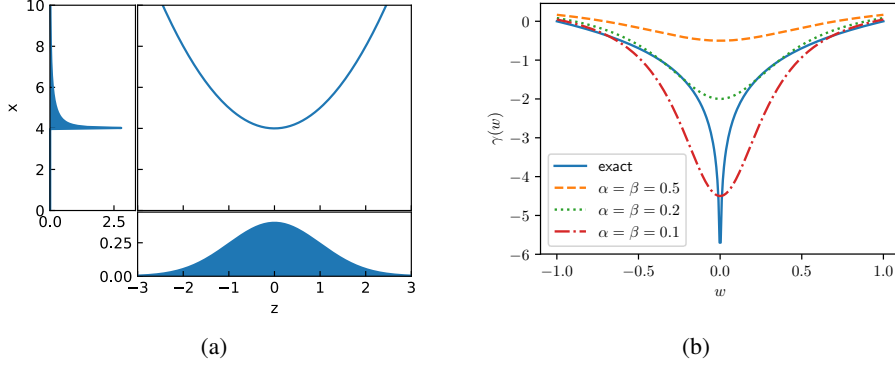


Figure 1: (a) An example of a univariate distribution under a non-invertible transformation. The x-axis shows the initial distribution $z \sim \mathcal{N}(0, 1)$ and the transformation is a quadratic function $x = z^2 + 4$. The resulting distribution is visualized along the y-axis. (b) The comparison of the "regularization" term in the exact PDF and our variational approximation under a linear transformation $x = wz$. The x-axis shows w and the y-axis shows the value of the "regularization" term. The solid curve indicates the exact PDF and the other curves indicates our variational approximation with different α and β .

PDF of the resulting distribution is,

$$q(x) = \begin{cases} 0 & x < 4, \\ \frac{q(z=-\sqrt{x-4})}{|f'(-\sqrt{x-4})|} + \frac{q(z=\sqrt{x-4})}{|f'(\sqrt{x-4})|} & x > 4. \end{cases}$$

As shown in the example, the PDF is zero when no real roots exist and the PDF is a sum of multiple terms, one for each real root, in which each term takes the same form as the PDF under an invertible transformation in (5).

For multivariate distributions, a similar result could also be obtained. When the dimensionality of the initial variable D_z is bigger than the dimensionality of the target variable D_x , using the co-area formula from geometric measure theory [Federer, 1969], a similar formula can be derived for multivariate distributions,

$$q(\mathbf{x}) = \int \delta(\mathbf{x} - f(\mathbf{z})) g(\mathbf{z}) d\mathbf{z} = \int_{f^{-1}(\mathbf{x})} \frac{g(\mathbf{z})}{|\mathbf{J}_z^\top \mathbf{J}_z|^{\frac{1}{2}}} d\mathbf{z}, \quad (7)$$

where \mathbf{J}_z is the Jacobian of the function f at \mathbf{z} and the integral is over $f^{-1}(\mathbf{x})$, the surface defined by $f(\mathbf{z}) = \mathbf{x}$. Note that, the denominator in the multidimensional formula is different from the one in the univariate case, which is due to the fact that the Jacobian of a multivariate transformation may not be a square matrix. If x is a scalar, the PDF becomes a special case,

$$q(x) = \int_{f^{-1}(x)} \frac{g(\mathbf{z})}{|\nabla f|} d\mathbf{z}, \quad (8)$$

where the denominator is the norm of the gradient of the function f . If the dimensionality of the initial variable D_z is smaller than the dimensionality of the resulting variable D_x , the PDF is no longer finite, as it is a degenerate distribution.

3 Variational Non-invertible Flow

Although the PDF of a random variable generated from a non-invertible transformation can be formulated analytically, it depends on all the real roots of the transformation equation, which is non-trivial for complex transformation functions. This is because the conditional distribution $q(\mathbf{x}|\mathbf{z})$ is degenerate. To calculate the PDF at a given \mathbf{x} , it needs to find the corresponding values of \mathbf{z} that map to \mathbf{x} , which is formulated as the roots of the transformation equation. To avoid the root finding, we approximate the degenerate conditional distribution $q(\mathbf{x}|\mathbf{z})$ with a normal distribution with a small variance, which can be viewed as an approximation to the Dirac delta distribution. The Dirac Delta function can be define as the limit of the Gaussian PDF as the variance approaches zero, *i.e.*,

147 $\delta(\mathbf{x}) = \lim_{\alpha \rightarrow 0} \mathcal{N}(\mathbf{x}|0, \alpha \mathbf{I})$. We avoid the infinite probability density problem by setting α to be a
 148 small positive number, i.e.,

$$q(\mathbf{x}|\mathbf{z}) = \delta(\mathbf{x} - f(\mathbf{z})) = \lim_{\alpha \rightarrow 0} \mathcal{N}(\mathbf{x}|f(\mathbf{z}), \alpha \mathbf{I}) \approx \mathcal{N}(\mathbf{x}|f(\mathbf{z}), \alpha \mathbf{I}). \quad (9)$$

149 As a result, $q(\mathbf{x}|\mathbf{z})$ gives a small probability mass around \mathbf{x} , which avoids $p(\mathbf{x})$ to be degenerate if
 150 $D_x > D_z$. We no longer have an analytical expression for $q(\mathbf{x})$, which needs to be estimated from
 151 the intractable marginal distribution $q(\mathbf{x}) = \int q(\mathbf{x}|\mathbf{z})q(\mathbf{z})d\mathbf{z}$.

152 3.1 Variational lower bound with auxiliary distribution

153 As $q(\mathbf{x})$ is no longer closed form, we cannot directly use it as a variational posterior for VI. To
 154 avoid this problem, we introduce an auxiliary distribution $\tilde{q}(\mathbf{z}|\mathbf{x})$. For simplicity, we use a normal
 155 distribution $\tilde{q}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\tilde{f}(\mathbf{x}), \beta \mathbf{I})$, where \tilde{f} is a function represented by a DNN, but other
 156 distributions are also applicable. With the auxiliary distribution, we define an unbiased estimator of
 157 the marginal likelihood $p(\mathbf{y})$ by drawing Monte Carlo samples,

$$p(\mathbf{y}) = \int q(\mathbf{x}, \mathbf{z}) \frac{p(\mathbf{y}, \mathbf{x}) \tilde{q}(\mathbf{z}|\mathbf{x})}{q(\mathbf{x}, \mathbf{z})} d\mathbf{x} d\mathbf{z} \approx \frac{1}{N} \sum_{i=1}^N \frac{p(\mathbf{y}, \mathbf{x}_i) \tilde{q}(\mathbf{z}_i|\mathbf{x}_i)}{q(\mathbf{x}_i, \mathbf{z}_i)}, \quad \mathbf{x}_i, \mathbf{z}_i \sim q(\mathbf{x}, \mathbf{z}), \quad (10)$$

158 where N is the number of drawn samples and $q(\mathbf{x}, \mathbf{z}) = q(\mathbf{x}|\mathbf{z})q(\mathbf{z})$. Using the Jensen's inequality, a
 159 variational lower bound can be derived for the log marginal likelihood,

$$\log p(\mathbf{y}) = \log \int q(\mathbf{x}, \mathbf{z}) \frac{p(\mathbf{y}, \mathbf{x}) \tilde{q}(\mathbf{z}|\mathbf{x})}{q(\mathbf{x}, \mathbf{z})} d\mathbf{x} d\mathbf{z} \geq \int q(\mathbf{x}, \mathbf{z}) \log \frac{p(\mathbf{y}, \mathbf{x}) \tilde{q}(\mathbf{z}|\mathbf{x})}{q(\mathbf{x}, \mathbf{z})} d\mathbf{x} d\mathbf{z} = \mathcal{L} \quad (11)$$

160 We denote the above lower bound as \mathcal{L} . Note that, although with an extra term in the nominator, it is
 161 still a lower bound of the log marginal likelihood. The relation to the usual variational lower bound
 162 can be seen by re-arranging the bound as follow:

$$\mathcal{L} = \int q(\mathbf{x}) \log \frac{p(\mathbf{y}, \mathbf{x})}{q(\mathbf{x})} d\mathbf{x} d\mathbf{z} - \langle \text{KL}(\tilde{q}(\mathbf{z}|\mathbf{x}) \| q(\mathbf{z}|\mathbf{x})) \rangle_{q(\mathbf{x})}, \quad (12)$$

163 where $q(\mathbf{x}) = \int q(\mathbf{x}|\mathbf{z})q(\mathbf{z})d\mathbf{z}$ is the marginal variational posterior and $q(\mathbf{z}|\mathbf{x})$ is the true "posterior"
 164 of the variational distribution $q(\mathbf{x}, \mathbf{z})$. The first term is the usual variational lower bound. As the KL
 165 divergence is greater than or equal to zero, \mathcal{L} is a lower bound of the first term and they become
 166 equal only if the auxiliary distribution $\tilde{q}(\mathbf{z}|\mathbf{x})$ is the same as the true variational posterior distribution
 167 $q(\mathbf{z}|\mathbf{x})$ for all \mathbf{x} .

168 3.2 Connection to normalizing flow

169 We approximate the lower bound of NFW with Monte Carlo samples,

$$\mathcal{L} \approx \frac{1}{K} \sum_{i=1}^K \left(\log \frac{p(\mathbf{y}, \mathbf{x}_i)}{q(\mathbf{z}_i)} + \log \frac{\tilde{q}(\mathbf{z}_i|\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{z}_i)} \right), \quad \mathbf{z}_i \sim q(\mathbf{z}), \quad (13)$$

170 where $\mathbf{x}_i = f(\mathbf{z}_i) + \alpha^{\frac{1}{2}} \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \mathbf{I})$. Note the first term is the same as the one for NF, which is

$$\mathcal{L} \approx \frac{1}{K} \sum_{i=1}^K \left(\log \frac{p(\mathbf{y}, \mathbf{x}_i)|_{\mathbf{x}_i=f(\mathbf{z}_i)}}{q(\mathbf{z}_i)} + \log |\mathbf{J}_{\mathbf{z}_i}| \right). \quad (14)$$

171 The first term is in favor of the location where the likelihood is high. Optimizing with this term alone
 172 will lead to the variational distribution that is a Dirac Delta distribution at the mode of $p(\mathbf{x}|\mathbf{y})$. The
 173 second term of (14) gives a regularization effect by encouraging the gradient of the transformation to
 174 be higher. The transformation with a higher gradient tends to spread the probability mass wider in the
 175 resulting distribution, which gives an opposite effect comparing to the first term. The second term of
 176 (13) gives a similar regularization effect. The denominator of the second term $q(\mathbf{x}|\mathbf{z})$ is independent
 177 of \mathbf{z} and \mathbf{x} and only depends on ϵ , which can be omitted in our analysis. According to the definition
 178 of $\tilde{q}(\mathbf{z}_i|\mathbf{x}_i)$, the nominator of the second term is,

$$\log \tilde{q}(\mathbf{z}_i|\mathbf{x}_i) = -\frac{Q}{2} \log 2\pi\beta - \frac{1}{2\beta} \left\| \mathbf{z}_i - \tilde{f}(f(\mathbf{z}_i) + \alpha^{\frac{1}{2}} \epsilon_i) \right\|^2. \quad (15)$$

It is similar to denoising autoencoder [Vincent et al., 2008], where the main difference is that the noise is added after the first transformation instead of to the input. This term measures that how well \mathbf{z} can be recovered from the transformed distribution $f(\mathbf{z})$ under noise corruption. Intuitively, to minimize the effect of noise corruption, this term encourages the transformation f to spread the probability mass as wide as possible. Conceptually, this produces a similar regularization effect as the second term of (14). Fig. 1b shows a quantitative comparison of the second terms in NF and NFW given a 1D linear transformation. As shown in the figure, both terms are concave with the same minimal. The smaller α and β is, the steeper the regularization effect is in our variational lower bound. See the Appendix for more details of this comparison.

4 Related Works

Normalizing flow [Jimenez Rezende and Mohamed, 2015] use invertible transformations to construct complex distributions from simple ones. Follow-ups works [Kingma et al., 2016, Dinh et al., 2017, Huang et al., 2018] addresses the issue about expensive computation by constructing transformation functions with a triangular Jacobians.

Non-invertible transformations have used in implicit generative models [Mohamed and Lakshminarayanan, 2016, Nowozin et al., 2016, Huszár, 2017, Tran et al., 2017, Li and Turner, 2018, Mescheder et al., 2017, Shi et al., 2018, Lueckmann et al., 2018]. A major challenge of implicit generative models is the density ratio estimation in the objective, which is often implemented with an adversarial networks. Semi-implicit variational inference [Yin and Zhou, 2018, Titsias and Ruiz, 2019, Molchanov et al., 2019] is a hybrid approach which uses a non-invertible transformation and assumes the target random variable follows a reparameterizable distribution depended on the outcome of the transformation. Although it avoids the problem of intractable PDF, the PDF of the target variable becomes a marginal distribution. Yin and Zhou [2018], Molchanov et al. [2019] use an upper bound of the PDF of the resulting marginal distribution, which is only tight if the number of samples is infinite. In practice, this requires a high number of samples. Titsias and Ruiz [2019] derives an unbiased estimator of the log PDF of the marginal distribution, which uses a Markov Chain Monte Carlo (MCMC) method to draw samples from the true posterior distribution of the transformed random variables. Our problem formulation is very related to SIVI, but differs mainly from the two aspects: (1) using an auxiliary distribution to derive a tight variational lower bound (2) our method emphasizes on the non-invertible transformation, uses the conditional distribution of the target variable only as an approximation to the Dirac delta distribution.

The idea of using latent variables in variational posterior starts with using mixtures [Bishop et al., 1998, Gershman et al., 2012, Salimans and Knowles, 2013, Guo et al., 2016, Miller et al., 2017] and later extends with hierarchical models [Tran et al., 2015, Ranganath et al., 2016, Maaløe et al., 2016]. To resolve the issue about the marginal PDF of variational posterior, Agakov and Barber [2004], Ranganath et al. [2016], Tran et al. [2015], Maaløe et al. [2016] exploit an auxiliary variable / distribution to derive a variational lower bound.

5 Experiments

We demonstrate the performance of NFW as a variational posterior in several tasks.

Synthetic distributions. We first demonstrate the capability of NFW in terms of capturing complex distributions. We use NFW to match three synthetic distribution by minimizing the KL divergence between the variational distribution and the synthetic distribution $\text{KL}(q(\mathbf{x}) \| p(\mathbf{x}))$. We take the synthetic distribution used in [Yin and Zhou, 2018, Titsias and Ruiz, 2019]. For all the synthetic distributions, we use a 2D normal distribution $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as the initial distribution and use a DNN with two hidden layers with 50 hidden units and the rectified linear (ReLU) unit as the transformation function. The auxiliary distribution uses a DNN with the mirrored architecture as the transformation function. We randomly initialized the weights of all the DNNs and initialize $\alpha = \sigma(-7)$ and $\beta = \sigma(-5)$, where $\sigma(a) = \log(1 + \exp(a))$ is the soft-plus function. The number of samples used in learning is 1000. We optimize the lower bound using Adam [Kingma and Ba, 2015] with the learning rate 0.001 for the 100k iterations. The contour plots of the synthetic distributions and 300 samples drawn from the fitted variational distributions are shown in Fig. 2. The samples from the fitted distributions match well with all the synthetic distribution. Note that, although the auxiliary

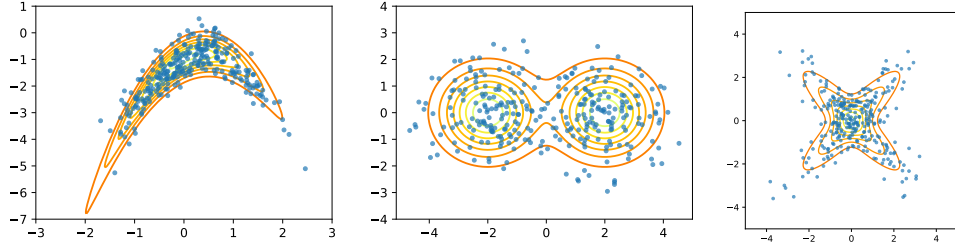


Figure 2: The contour plots of the synthetic distributions and the samples drawn from the variational distributions matching with the synthetic distributions using NFW.

method	MNIST	Fashion-MNIST
NFW	-91.24	-116.89
meanfield	-98.29	-126.73
[Yin and Zhou, 2018]	-97.77	-121.53
[Titsias and Ruiz, 2019]	-94.09	-110.72

Table 1: Test log-likelihood of VAE on binarized MNIST and binarized Fashion-MNIST

distribution of NFW encourages a one-to-one mapping between \mathbf{z} and \mathbf{x} , it does not prevent NFW to generate multi-modal distributions.

Variational autoencoder. Latent variable models such as variational autoencoder (VAE) [Kingma and Welling, 2014] model the generation process from a latent representation \mathbf{x} to an observed data point \mathbf{y} as a conditional distribution $p(\mathbf{y}|\mathbf{x})$. With a typically un-informative prior distribution $p(\mathbf{x})$, the goal is to learn the parameters of the conditional distribution $p(\mathbf{y}|\mathbf{x})$, while estimating the posterior distribution $p(\mathbf{x}|\mathbf{y})$. Such model especially VAE becomes a testbed for many variational inference methods, as learning a good generative model $p(\mathbf{y}|\mathbf{x})$ requires a good mechanism for estimating variational posteriors $p(\mathbf{x}|\mathbf{y})$.

We compare with previous works by using NFW to represent the variational posterior of VAE. As VAE uses amortized inference, the variational posterior $q(\mathbf{x}|\mathbf{y})$ depends on the observed data representation \mathbf{y} . When applying NFW, we concatenate the data representation \mathbf{y} with the initial random variable \mathbf{z} and use a DNN as the transformation function that takes the concatenated representation as the input. The initial random variable \mathbf{z} follows a normal distribution of which the dimensionality is the same as \mathbf{x} . The DNN has two hidden layers with ReLU. The DNN in the auxiliary distribution has the mirrored architecture. We initialize $\alpha = \sigma(-5)$ and $\beta = \sigma(-3)$.

We train VAE on two standard datasets: the stochastically binarized MNIST [Salakhutdinov and Murray, 2008] and the Fashion-MNIST [Xiao et al., 2017]. Both datasets have 60,000 images of 28x28 resolution for training and 10,000 images for testing. We follow [Titsias and Ruiz, 2019] to binarize the Fashion-MNIST images by thresholding at 0.5. In VAE, we model binary data by applying a Sigmoid transformation to the output of DNN and use the outcome as the probability of Bernoulli distribution. The dimensionality of \mathbf{x} is ten. The DNN in VAE $p(\mathbf{y}|\mathbf{x})$ has two hidden layers with 200 hidden units and ReLU non-linearity. The performance of a model is evaluated as the average marginal likelihood log-likelihood of the VAE on the set of test images. The marginal log-likelihood is estimated via importance sampling as in (10). This estimator has lower variance compares with the importance sampling approach used in [Yin and Zhou, 2018, Titsias and Ruiz, 2019]. The results of NFW are compared with previous methods in Tab. 1.

Imputation with VAE. After training a VAE, we demonstrate NFW by inferring the posterior distribution of VAE when the image is only partially observed. When only observing part of a handwritten digit, there are often multiple possibilities to complete the digit. A good posterior captures as many possibilities as possible. The requires the formulation of the posterior distribution to be very flexible. To set up such an experiment, we randomly select one image per digit and binarize the images by sampling from the Bernoulli distributions that take the normalized intensities as the probabilities of being one. We infer a variational posterior distribution for each image by only observing the top 12 rows of the binarized images, which are shown in the first column in Fig. 3. We compare NFW and mean field on this task using the same VAE. For NFW, we use a 10D normal

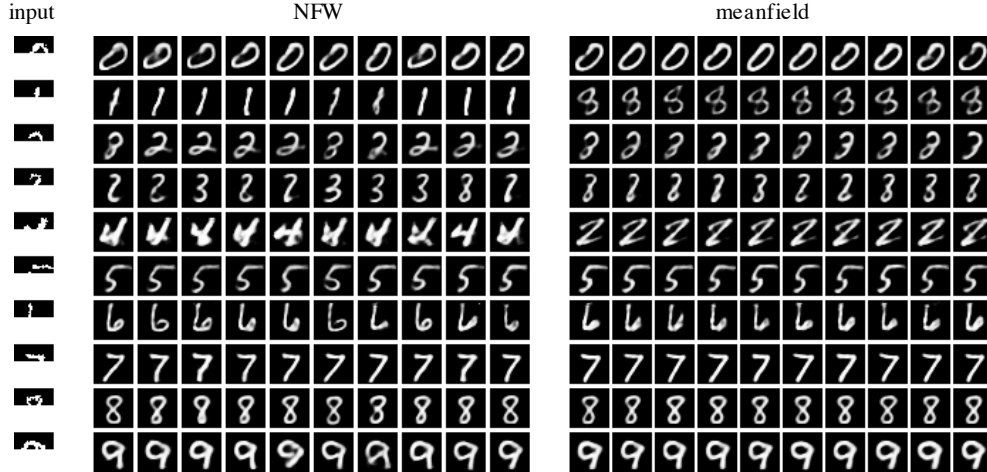


Figure 3: The image samples from the variational posteriors that are inferred from partially observed images. The partially observed images are shown in the 1st column. The samples from NFW are shown in the middle column and the samples from mean field are shown in the 3rd column.

dataset	NFW	meanfield	PBP	Dropout
kin8nm	1.19 (0.03)	1.16 (0.03)	0.90 (0.01)	0.95 (0.01)
power	-2.78 (0.04)	-2.79 (0.03)	-2.84 (0.01)	-2.80 (0.01)
energy	-1.32 (0.21)	-1.59 (0.11)	-2.04 (0.02)	-1.99 (0.02)
concrete	-3.05 (0.19)	-3.07 (0.16)	-3.16 (0.02)	-3.04 (0.02)

Table 2: The average test log-likelihood of BNN on UCI regression benchmarks. The shown results are the average of 20 random partitions and the standard deviations are shown in parentheses.

distribution as the initial distribution and use a two hidden layer DNN with 300 hidden units as the transformation function. The setting of optimization is the same as for training the VAE. For each image, we randomly draw 10 samples from the inferred variational posterior and visualize in Fig. 3, where the middle column is the samples from NFW and the right column is the samples from the mean field posterior. The samples from NFW are more diverse, as NFW provides a more flexible posterior representation.

BNN regression. A major challenge of using NF for BNN is the dimensionality of the variational distribution. With a non-invertible transformation, we can easily construct a correlated high dimensional variational posterior. We use NFW to construct the variational posterior for BNN and evaluated on a few UCI regression benchmarks. We follow the common protocol of using 90% for training and 10% for testing. The experiments are repeated 20 times with different random partitions. For all the regression datasets and runs, we use the same setting. The BNN has one hidden layer with 50 units and ReLU. The dimensionality of the initial distribution is 100, the transformation function is a DNN with two hidden layers (300-500 units) and ReLU. We optimize with Adam at the learning rate of 0.01 for 1000k iterations and 0.001 for another 1000k iterations. The number of samples is 10. We initialize $\alpha = \sigma(-5)$ and $\beta = \sigma(-3)$. We compare our method with the meanfield variational posterior with the reparameterization trick. The results are shown in Tab. 2. The results of NFW and meanfield use the same set of random partitions. The results of PBP [Hernández-Lobato and Adams, 2015] and Dropout [Gal and Ghahramani, 2016] are directly taken from the papers.

6 Conclusion

We show that the PDF of the resulting distribution that is transformed with a non-invertible function can be derived analytically. To address the issue about computing all the real roots of the transformation equation, we derive a variational lower bound by introducing an auxiliary distribution. We demonstrate that the auxiliary distribution in NFW produces a similar regularization effect as the log determinant in NF when the transformation is invertible. This provides some insights about the connection between the NF-based methods and the SIVI methods.

References

- Felix V Agakov and David Barber. An auxiliary variational method. In *International Conference on Neural Information Processing*, 2004.
- George Arfken. *Mathematical Methods for Physicists*. Academic Press, Inc., 1985.
- Christopher M Bishop, Neil D Lawrence, Tommi Jaakkola, and Michael I Jordan. Approximating posterior distributions in belief networks using mixtures. In *Advances in neural information processing systems*, 1998.
- Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.
- Herbert Federer. *Geometric measure theory*. Springer-Verlag New York Inc., 1969.
- Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. In *Advances in Neural Information Processing Systems 31*, 2018.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.
- Samuel Gershman, Matt Hoffman, and David Blei. Nonparametric variational inference. In *International Conference on Machine Learning*, 2012.
- Fangjian Guo, Xiangyu Wang, Kai Fan, Tamara Broderick, and David B Dunson. Boosting variational inference. *arXiv preprint arXiv:1611.05559*, 2016.
- José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, 2015.
- Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, 2018.
- Ferenc Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.
- Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 2015.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.
- Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.
- Yingzhen Li and Richard E Turner. Gradient estimators for implicit models. In *International Conference on Learning Representations*, 2018.
- Jan-Matthis Lueckmann, Giacomo Bassetto, Theofanis Karaletsos, and Jakob H Macke. Likelihood-free inference with emulator networks. *arXiv preprint arXiv:1805.09294*, 2018.
- Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. In *International Conference on Machine Learning*, 2016.
- Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, 2017.
- Andrew C Miller, Nicholas J Foti, and Ryan P Adams. Variational boosting: Iteratively refining posterior approximations. In *International Conference on Machine Learning*, 2017.

337 Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv*
338 *preprint arXiv:1610.03483*, 2016.

339 Dmitry Molchanov, Valery Kharitonov, Artem Sobolev, and Dmitry Vetrov. Doubly semi-implicit
340 variational inference. In *International Conference on Artificial Intelligence and Statistics*, 2019.

341 Christian A Naesseth, Francisco JR Ruiz, Scott W Linderman, and David M Blei. Reparameterization
342 gradients through acceptance-rejection sampling algorithms. In *International Conference on*
343 *Artificial Intelligence and Statistics*, 2016.

344 Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers
345 using variational divergence minimization. In *Advances in neural information processing systems*,
346 2016.

347 Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International*
348 *Conference on Machine Learning*, 2016.

349 Francisco R Ruiz, Michalis Titsias RC AUEB, and David Blei. The generalized reparameterization
350 gradient. In *Advances in neural information processing systems*, 2016.

351 Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In
352 *Proceedings of the 25th international conference on Machine learning*, 2008.

353 Tim Salimans and David Knowles. Fixed-form variational posterior approximation through stochastic
354 linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

355 Jiaxin Shi, Shengyang Sun, and Jun Zhu. Kernel implicit variational inference. In *International*
356 *Conference on Learning Representations*, 2018.

357 Michalis K Titsias and Francisco JR Ruiz. Unbiased implicit variational inference. In *International*
358 *Conference on Artificial Intelligence and Statistics*, 2019.

359 Dustin Tran, Rajesh Ranganath, and David M Blei. The variational gaussian process. In *International*
360 *Conference on Learning Representations*, 2015.

361 Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free
362 variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533,
363 2017.

364 Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and
365 composing robust features with denoising autoencoders. In *Proceedings of the 25th international*
366 *conference on Machine learning*, 2008.

367 Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking
368 machine learning algorithms, 2017.

369 Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. In *International Conference*
370 *on Machine Learning*, 2018.