Variational Inference with Orthogonal Normalizing Flows

Leonard Hasenclever
University of Oxford

Jakub M. Tomczak
University of Amsterdam

Rianne van den Berg
University of Amsterdam

Max Welling
University of Amsterdam

1 Introduction

Normalizing flows Variational inference relies on flexible approximate posterior distributions. In many settings very simple posteriors such as diagonal covariance Gaussians are used. Rezende and Mohamed [2015] propose a way to construct more flexible posteriors by transforming a simple base distribution with a series of invertible transformations with easily computable Jacobians. The resulting transformed density after one such transformation is given by:

$$p_1(\mathbf{z}') = p_0(\mathbf{z}) \left| \det \left(\frac{\partial f(\mathbf{z})}{\partial \mathbf{z}} \right) \right|^{-1}, \tag{1}$$

where $\mathbf{z}' = f(\mathbf{z})$, $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^D$ and $f : \mathbb{R}^D \mapsto \mathbb{R}^D$ is an invertible function. While in general the cost of computing the Jacobian will be $\mathcal{O}(D^3)$, for practical use it is desirable to design transformations with more efficiently computable Jacobians.

Planar flow Rezende and Mohamed [2015] propose two families of parametrized normalizing flows that fulfill these criteria: planar and radial flows. Planar flows are given by the following transformation:

$$\mathbf{z}' = \mathbf{z} + \mathbf{u} h(\mathbf{w}^T \mathbf{z} + b), \tag{2}$$

where $\mathbf{u}, \mathbf{w} \in \mathbb{R}^D$, $b \in \mathbb{R}$ and h is a suitable smooth activation function. Rezende and Mohamed [2015] show that given the choice of activation function $h = \tanh$, transformations of this kind are invertible as long as $\mathbf{u}^T \mathbf{w} \geq -1$. By the *Matrix Determinant Lemma* the Jacobian of this transformation is given by:

$$\det \frac{\partial \mathbf{z}'}{\partial \mathbf{z}} = \det \left(\mathbf{I} + \mathbf{u} h'(\mathbf{w}^T \mathbf{z} + b) \mathbf{w}^T \right) = 1 + \mathbf{u}^T h'(\mathbf{w}^T \mathbf{z} + b) \mathbf{w}. \tag{3}$$

This lemma is a special case of Sylvester's determinant identity:

Theorem 1 (Sylvester's determinant identity)**.** *For all* $\mathbf{A} \in \mathbb{R}^{M \times D}, \mathbf{B} \in \mathbb{R}^{D \times M}$*,*

$$\det \left(\mathbf{I}_M + \mathbf{A}\mathbf{B} \right) = \det \left(\mathbf{I}_D + \mathbf{B}\mathbf{A} \right), \tag{4}$$

where $\mathbf{I}_M$ *and* $\mathbf{I}_D$ *are* M *and* D*-dimensional identity matrices, respectively.*

After applying K flows, the final latent stochastic variables are given by $\mathbf{z}_K = f_K \circ \ldots f_2 \circ f_1(\mathbf{z}_0)$, with the initial simple posterior density for $\mathbf{z}_0$ given by a diagonal Gaussian $\mathcal{N}(\mathbf{z}_0 | \boldsymbol{\mu}, \sigma^2 \mathbf{I})$. In the amortized variational inference model of Rezende and Mohamed [2015], both $\boldsymbol{\mu}$ and σ, as well as all of the flow parameters $\{\mathbf{u}^{(k)}, \mathbf{w}^{(k)}, b^{(k)}\}_{k=1}^K$ are produced by the outputs of a deep neural inference network that maps an input vector $\mathbf{x}$ to these parameters.

Second workshop on Bayesian Deep Learning (NIPS 2017), Long Beach, CA, USA.

Paper contribution In this paper we use Sylvester's determinant identity to introduce a new normalizing flow. In order to guarantee invertible flows and to ensure efficient computation of the Jacobian determinant we use orthogonal weight matrices in the flow. We refer to the resulting flow as the *orthogonal normalizing flow*.

2 Orthogonal Normalizing Flows

In practice, many planar flow transformations are required to transform a simple base distribution into a flexible distribution, especially for high dimensional latent spaces. In addition, planar flows tend to be hard to train and sensitive to initializations. Kingma et al. [2016] argue that is this due to the fact that the activation function effectively acts as a single-unit MLP. This raises the question if we can construct more powerful invertible transformations with easily computable Jacobians.

Let us consider the following transformation:

$$\mathbf{z}' = \mathbf{z} + \mathbf{A}h(\mathbf{B}\mathbf{z} + \mathbf{b}), \tag{5}$$

where $\mathbf{A} \in \mathbb{R}^{D\times M}, \mathbf{B} \in \mathbb{R}^{M\times D}$ and $\mathbf{b} \in \mathbb{R}^M$. This transformation has Jacobian determinant:

$$\det \frac{\partial \mathbf{z}'}{\partial \mathbf{z}} = \det \left(\mathbf{I}_M + \operatorname{diag}\left(h'(\mathbf{B}\mathbf{z} + \mathbf{b})\right)\mathbf{B}\mathbf{A}\right), \tag{6}$$

which follows from Sylvester's determinant identity. This trick reduces the determinant of a $D \times D$ matrix to a determinant of an $M \times M$ matrix. However, in general this transformation will not necessarily be invertible. Therefore, we propose the following special case of the above transformation:

$$\mathbf{z}' = \mathbf{z} + \mathbf{W}\mathbf{D}h(\tilde{\mathbf{D}}\mathbf{W}^T\mathbf{z} + \mathbf{b}) = \phi(\mathbf{z}), \tag{7}$$

where $\mathbf{D}$ and $\tilde{\mathbf{D}}$ are diagonal matrices, and $\mathbf{W} = (\mathbf{w}_1 \dots \mathbf{w}_M)$ with the columns $\mathbf{w}_m$ forming an orthonormal set. By theorem 1 the determinant of the Jacobian of this transformation reduces to:

$$\det \frac{\partial \mathbf{z}'}{\partial \mathbf{z}} = \det \left(\mathbf{I}_M + \operatorname{diag}\left(h'(\tilde{\mathbf{D}}\mathbf{W}^T\mathbf{z} + \mathbf{b})\right)\tilde{\mathbf{D}}\mathbf{W}^T\mathbf{W}\mathbf{D}\right) \tag{8}$$

$$= \det \left(\mathbf{I}_M + \operatorname{diag}\left(h'(\tilde{\mathbf{D}}\mathbf{W}^T\mathbf{z} + \mathbf{b})\right)\tilde{\mathbf{D}}\mathbf{D}\right), \tag{9}$$

which is efficient to compute. The following theorem gives a sufficient condition for this transformation to be invertible.

Theorem 2. *Let $h : \mathbb{R} \longrightarrow \mathbb{R}$ be a smooth function with bounded, positive derivative. Then, if the diagonal entries of $\mathbf{D}$ and $\tilde{\mathbf{D}}$ satisfy $d_i\tilde{d}_i > -1/\|h'\|_\infty$ the transformation given by* (7) *is invertible.*

Proof. Recall that one-dimensional real functions with strictly positive derivatives are invertible. Note that the columns of $\mathbf{W}$ are orthonormal, such that they span a (sub)space of $\mathbb{R}^D$. Let $\mathcal{W} = \text{span}\{\mathbf{w}_1, \dots, \mathbf{w}_M\}$ and $\mathcal{W}^\perp$ its orthogonal complement. We can decompose $\mathbf{z} = \mathbf{z}_\parallel + \mathbf{z}_\perp$, where $\mathbf{z}_\parallel \in \mathcal{W}$ and $\mathbf{z}_\perp \in \mathcal{W}^\perp$. Similarly we can decompose $\mathbf{z}' = \mathbf{z}'_\parallel + \mathbf{z}'_\perp$. Note that $\mathbf{W}\mathbf{W}^T\mathbf{z} = \mathbf{z}_\parallel$ and $\mathbf{W}^T\mathbf{z}_\perp = \mathbf{0}$ by definition. Clearly, $\mathbf{W}\mathbf{D}h(\tilde{\mathbf{D}}\mathbf{W}^T\mathbf{z} + \mathbf{b}) \in \mathcal{W}$, hence leading to the unique solution for the orthogonal component $\mathbf{z}_\perp = \phi(\mathbf{z})_\perp = \mathbf{z}'_\perp$. Thus, it suffices to consider the transformation in the directions of $\mathbf{w}_1, \dots, \mathbf{w}_m$. Multiplying (7) by $\mathbf{W}^T$ from the left gives:

$$\underbrace{\mathbf{W}^T\mathbf{z}'}_{\mathbf{v}'} = \underbrace{\mathbf{W}^T\mathbf{z}}_{\mathbf{v}} + \mathbf{D}h(\tilde{\mathbf{D}}\underbrace{\mathbf{W}^T\mathbf{z}}_{\mathbf{v}} + \mathbf{b}) = (f_1(v_1), \dots, f_m(v_m))^T, \tag{10}$$

where the vectors $\mathbf{v}$ and $\mathbf{v}'$ contain the respective coordinates of $\mathbf{z}_\parallel$ and $\mathbf{z}'_\parallel$ in the directions of $\mathbf{w}_1, \dots, \mathbf{w}_m$. The dimensions in (10) are completely independent and each dimension is transformed by a real function $f_i(v) = v + d_i h(\tilde{d}_i v + b_i)$. Consider a single dimension i of (10). Since $\|h'\|_\infty d_i\tilde{d}_i > -1$, we have $f'_i(v) > 0$ and thus f_i is invertible. In this transformation $\mathbf{z}_\perp$ is left unchanged while $\mathbf{z}_\parallel$ is transformed in an invertible way. Hence the whole transformation is invertible. □

For the case of $h(x) = \tanh(x)$, creating two diagonal matrices with $d_i\tilde{d}_i > -1/\|h'\|_\infty = -1$ can be achieved by taking two random diagonal matrices $\widehat{\mathbf{D}}^1$ and $\widehat{\mathbf{D}}^2$ and transforming them according to $\mathbf{D} = \tanh(\widehat{\mathbf{D}}^1)$ and $\tilde{\mathbf{D}} = \tanh(\widehat{\mathbf{D}}^2)$, since $-1 < \tanh(x) < 1\ \forall x \in \mathbb{R}$.

Last but not least, the flow in Eq. (7) relies on the parameter matrix $\mathbf{W}$ remaining orthogonal throughout the training process. We ensure this property by applying an iterative procedure proposed by Björck and Bowie [1971], Kovarik [1970]:

$$\mathbf{W}^{(k+1)} = \mathbf{W}^{(k)} \left(\mathbf{I} + \frac{1}{2} \left(\mathbf{I} - \mathbf{W}^{(k)\top} \mathbf{W}^{(k)} \right) \right). \tag{11}$$

with a sufficient condition for convergence given by $\|\mathbf{W}^{(0)\top}\mathbf{W}^{(0)} - \mathbf{I}\|_2 < 1$. Here the 2-norm of a matrix $\mathbf{X}$ refers to $\|\mathbf{X}\|_2 = \lambda_{\max}(\mathbf{X})$, with $\lambda_{\max}(\mathbf{X})$ representing the largest singular value of $\mathbf{X}$. In our experimental evaluations we ran the iterative procedure until $\|\mathbf{W}^{(k)\top}\mathbf{W}^{(k)} - \mathbf{I}\|_F \leq \epsilon$, with $\|\mathbf{X}\|_F$ the Frobenius norm, and ϵ a small convergence threshold. We observed that running this procedure up to 30 steps was sufficient to ensure convergence with respect to this threshold.

Since this orthogonalization procedure is differentiable, it allows for the calculation of gradients with respect to $\mathbf{W}^{(0)}$ by backpropagation, such that any standard optimization scheme such as stochastic gradient descent can be used for updating the flow weights.

In Figure 1, the effect of planar and orthogonal normalizing flows for a Gaussian initial density is shown. Note that since the latent space is only of size 2 for this visualization, at most $M = 2$ orthogonal vectors can appear in the orthogonal flow parameter matrix $\mathbf{W}$. We expect the orthogonal normalizing flow to be especially effective for higher dimensional latent spaces.

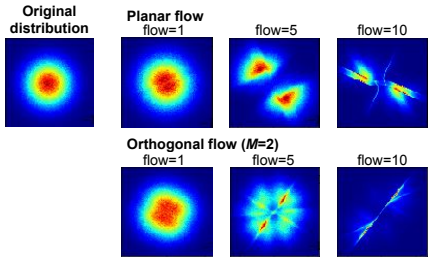

Figure 1: A transformation of a standard Gaussian posterior in 2D using the planar flow (*top*) and the orthogonal flow (*bottom*).

3 Results and discussion

In our experiments we compared the orthogonal flow with the planar flow on the statically binarized MNIST dataset. We utilized the normalizing flow in the Variational Autoencoder (VAE) [Kingma and Welling, 2013, Rezende et al., 2014] with two hidden fully connected layers (400 hidden units per layer) with softplus non-linearity in the encoder and the decoder and two simple output layers in the encoder giving $\boldsymbol{\mu}_0$ and $\boldsymbol{\sigma}_0$. The latent space of stochastic hidden units was set to 40. We used the standard division of the data into sets of $50,000$, $10,000$ and $10,000$ images for training, validation and testing, respectively. We apply warm-up for 100 epochs, and use the Adam optimizer [Kingma and Ba, 2014] for parameter optimization. The results are presented in Figure 2. The results clearly show that the model benefits both from more flows as well as a larger number of orthonormal vectors per flow.

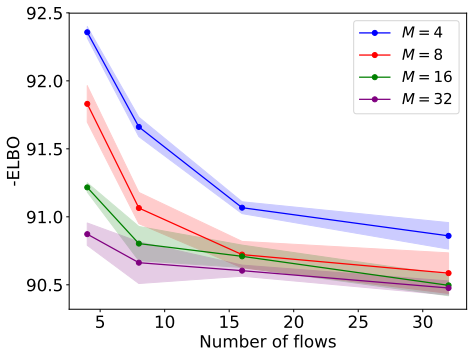

Figure 2: Results of the VAE with the newly proposed orthogonal flow in terms of the negative evidence lower bound (ELBO) on the test MNIST data. For each run the means are shown, with the shaded areas representing one standard deviation, computed with three runs for each point.

We used importance sampling [Rezende et al., 2014] to approximate the negative log-likelihood (NLL). The NLL for $K = 16$ orthogonal flows with $M = 16$ is equal to 85.9, compared to 86.5, 85.7 and 85.1 for 20, 40 and 80 planar flows, respectively.[1] This indicates that applying orthogonal flows leads to almost the same level of performance as the best performing model with planar flows while requiring four times fewer flows.

In the future, we plan to evaluate orthogonal normalizing flows more thoroughly as a tool in variational inference and explore other applications such as semi-supervised learning.

[1]The results for the planar flow are taken from [Rezende and Mohamed, 2015] where a different architecture than ours was used.

Acknowledgments

LH is funded by the UK EPSRC OxWaSP CDT through grant EP/L016710/1. JMT is funded by the European Commission within the MSC-IF (Grant No. 702666). RvdB is funded by SAP SE.

References

Åke Björck and Clazett Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *NIPS*, pages 4743–4751. 2016.

Zdislav Kovarik. Some iterative methods for improving orthonormality. *SIAM Journal on Numerical Analysis*, 7(3):386–389, 1970.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, pages 1530–1538, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.