Block Neural Autoregressive Flow

Nicola De Cao[1,2], Wilker Aziz[2], and Ivan Titov[1,2]
[1]University of Edinburgh, [2]University of Amsterdam

Key Ideas

- Neural autoregressive flow (NAF) by Huang et al. (2018)
  - PRO: **universal approximator of density functions**
  - CONS: **hyper-network → parameter num. grows quadratically**
- We propose **Block Neural Autoregressive Flow (B-NAF)**
  - a more compact **universal approximator of density functions,** directly modelled as a single **feed-forward network**
  - comparable in performance while using **orders of magnitude fewer parameters**

Introduction

A **normalising flows (NFs)** maps two density functions via a differentiable bijection (f):

$$p_Y(y) = p_X(x) \,|\det \mathbf{J}_{f(x)}|^{-1}$$

NFs are useful for learning densities: wide used in **density estimation** and **variational inference**

Usually, a density is decomposed in an **autoregressive** way:

$$p_X(x) = p_{X_1}(x_1) \prod_{i=2}^{d} p_{X_i|X_{<i}}(x_i \,|\, x_{<i})$$ to have a **tractable Jacobian!**

The NF is decomposed in: $y_i = f_\theta^{(i)}(x_{\leq i}) = t_\theta^{(i)}(x_i, c_\theta^{(i)}(x_{<i}))$

invertible **transformer** ↗ ↖ **conditioner**

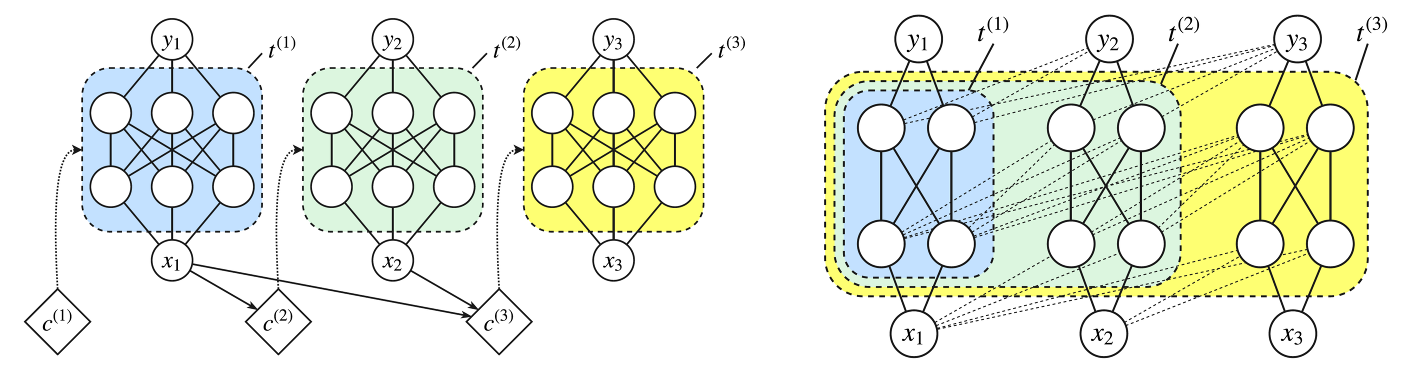
Invertibility depends on the transformers → **Trivially** invertible transformations **may not be expressive enough**

↓

Neural autoregressive flow (NAF) by Huang et al. (2018): **replaces hand-crafted transformers with invertible neural networks!**

The Jacobian is computed with **backpropagation**:

$$\mathbf{J}_{f_\theta(x)} = \left[\nabla_{h^{(l)}} y\right] \left[\nabla_{h^{(l-1)}} h^{(l)}\right] \dots \left[\nabla_x h^{(1)}\right]$$



(a) NAF: each $c(i)$ is a neural network that predicts pseudo-parameters for $t(i)$, which in turns processes $x(i)$.

(b) Our B-NAF: we do not use conditioner networks, instead we learn the flow network directly. Some weights are strictly positive (solid lines), others have no constraints (dashed lines).

Figure 1. Main differences between NAF (Huang et al., 2018) and our B-NAF.

Method

ADVANTAGES:
NAFs are **universal approximators of density functions**

DRAWBACKS:
NAFs are **hyper-networks** and therefore the number of **parameters scale quadratically!**

SOLUTION:
our model a universal approximator of density functions with single feed-forward network!

- we model each t directly as an NN without a conditioner
- we employ affine transformations with **positive weights** to process x_i ensuring **strict monotonicity and thus invertibility**

For each affine layer, the weight matrix W is a **lower-triangular block matrix** with **strictly positive diagonal blocks:**

$$W = \begin{bmatrix} \exp(B_{11}) & 0 & \dots & 0 \\ B_{21} & \exp(B_{22}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ B_{d1} & B_{d2} & \dots & \exp(B_{dd}) \end{bmatrix}$$

- **Universal approximator of densities:** we can arbitrarily increase the hidden layer dimension
- **Autoregressive:** lower triangular Jacobian
- **Stable:** the det-Jacobian can be computed in the **log-domain**
- **Efficient:** fewer parameters than NAF and easy-to-compute Jacobian

Model	POWER	GAS	HEPMASS	MINIBOONE	BSDS300
Real NVP	0.17	8.33	-18.71	-13.55	153.28
Glow	0.17	8.15	-18.92	-11.35	155.07
MADE	0.40	8.47	-15.15	-12.27	153.71
MAF	0.30	9.59	-17.39	-11.68	156.36
FFJORD	0.46	8.59	-14.92	-10.43	157.40
TAN	0.60	12.06	-13.78	-11.01	159.80
NAF-DDSF	0.62	11.96	-15.09	-8.86	157.43
Ours	**0.61**	**12.06**	**-14.71**	**-8.95**	**157.36**
Param. Gain	**2.29x**	**2.60x**	**17.94x**	**43.97x**	**8.24x**

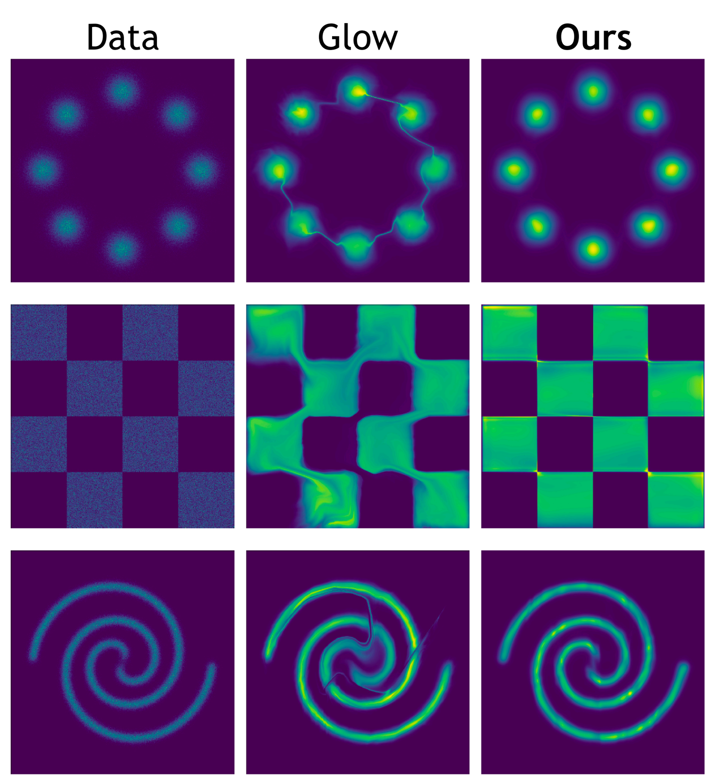
Table 1. Density estimation on 5 benchmark dataset. **B-NAF has comparable performance with NAF and order of magnitude fewer parameters.**

Results

Comparison with **Glow** (Kingma and Dhariwal, 2018) **on density estimation**

↓

discontinuities and **low-destiny** regions are better modelled by B-NAF

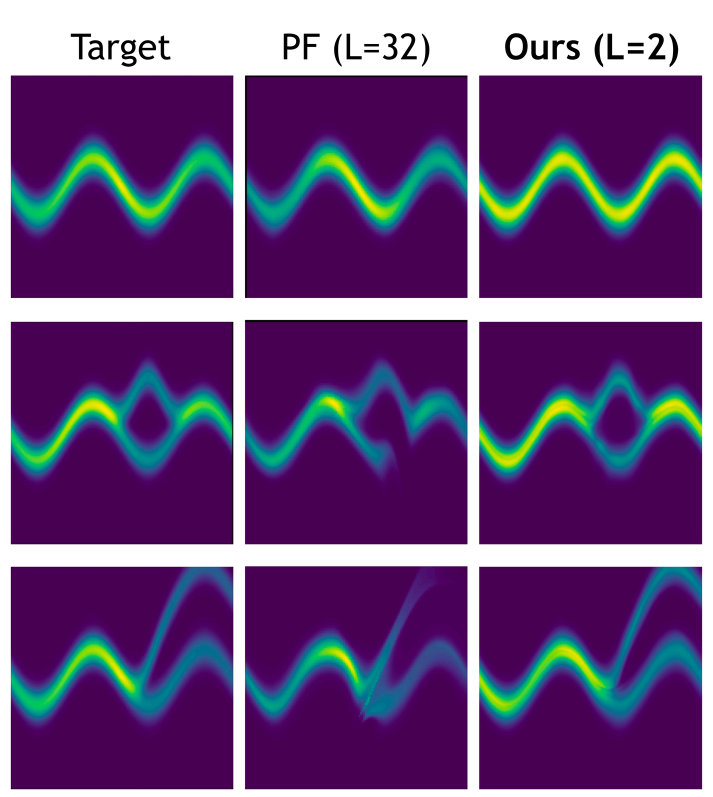


Comparison with **Planar Flows** (Rezendeand Mohamed, 2015) **on density matching**

2 layers of B-NAF **work better than** 32 layers of planar flows

More **shallow!**

Faster training!



Code available at https://github.com/nicola-decao/BNAF

Contact Information

Nicola De Cao
Ph.D. Candidate at University of Amsterdam
nicola.decao@gmail.com
https://nicola-decao.github.io
https://twitter.com/nicola_decao

References

- Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. (2018). Neural autoregressive flows. International Conference on Learning Representations.
- 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.
- Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. In Advances in Neural Information Processing Systems, pages 2338–2347.
- Oliva, J., Dubey, A., Zaheer, M., Poczos, B., Salakhutdinov, R., Xing, E., and Schneider, J. (2018). Transformation autoregressive networks. In Dy, J. and Krause, A., editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 3898–3907, Stockholmsmssan, Stockholm Sweden. PMLR.
- Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using real NVP. Proceedings of the 5th International Conference on Learning Representations (ICLR).
- Dua, D. and Karra Taniskidou, E. (2017). UCI machine learning repository.
- Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In Advances in Neural Information Processing Systems, pages 10236–10245.
- Larochelle, H. and Murray, I. (2011). The neural autoregressive distribution estimator. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 29–37.
- Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. In Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37, pages 1530–1538. JMLR. org.
- van den Berg, R., Hasenclever, L., Tomczak, J. M., and Welling, M. (2018). Sylvester normalizing flows for variational inference. 34th Conference on Uncertainty in Artificial Intelligence (UAI18).
- Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In Advances in neural information processing systems, pages 4743–4751.

Acknowledgements

We would like to thank George Papamakarios and Luca Falorsi for insightful discussions. This project is supported by SAP Innovation Center Network, ERC Starting Grant BroadSem (678254) and the Dutch Organization for Scientific Research (NWO) VIDI 639.022.518. Wilker Aziz is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825299 (Gourmet).