Primer on Normalizing Flows

Laurent Dinh

Google AI
Brain Team

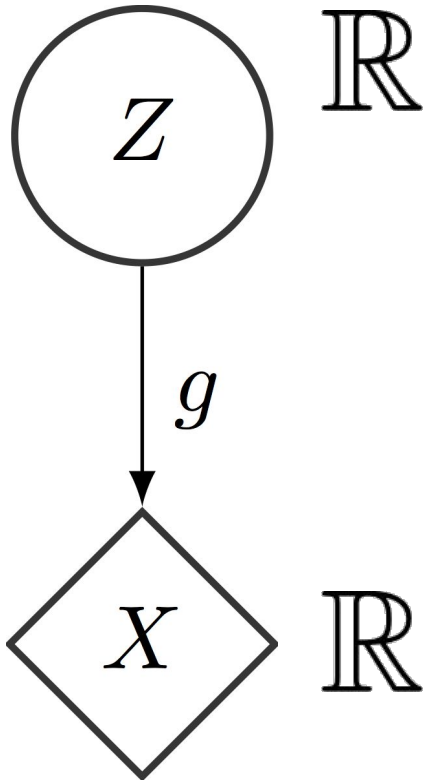
Motivation

Probability manipulation

- Density estimation / generative modelling
- Monte Carlo integration
  - Stochastic variational inference
  - Conditioning MCMC
  - …
- …

(Kingma & Dhariwal, 2018)

Change of variable formula



$$p_X(x)\,dx = p_Z(z)\left(\left|\frac{\partial z}{\partial x}\right|\right)dz$$

Density Volume

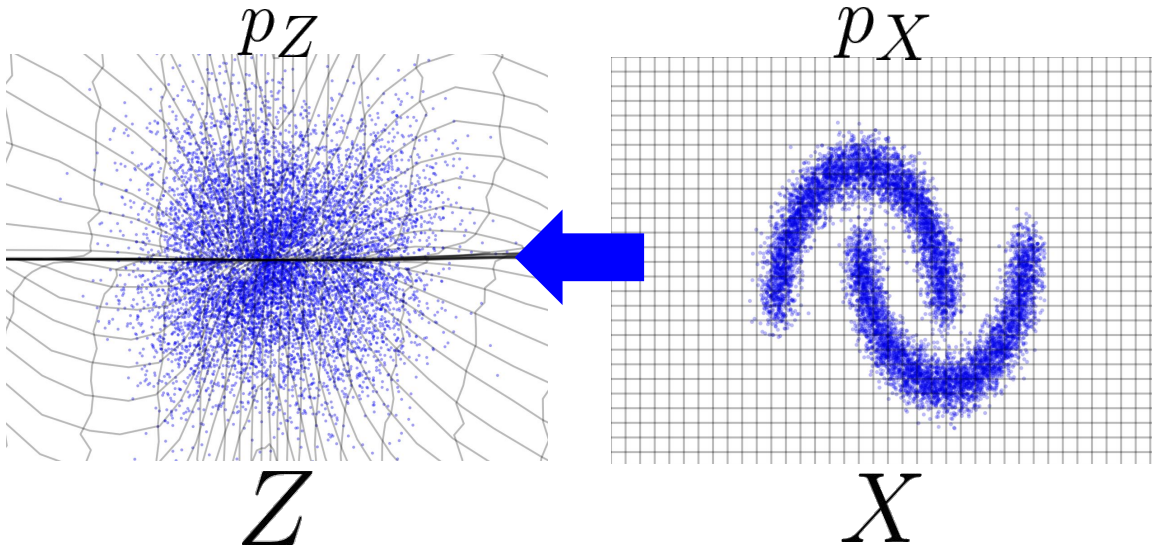
Mass

(Hyvärinen, 2000; Rippel & Adams, 2013; Dinh et al., 2014; Hoffman et al., 2019)

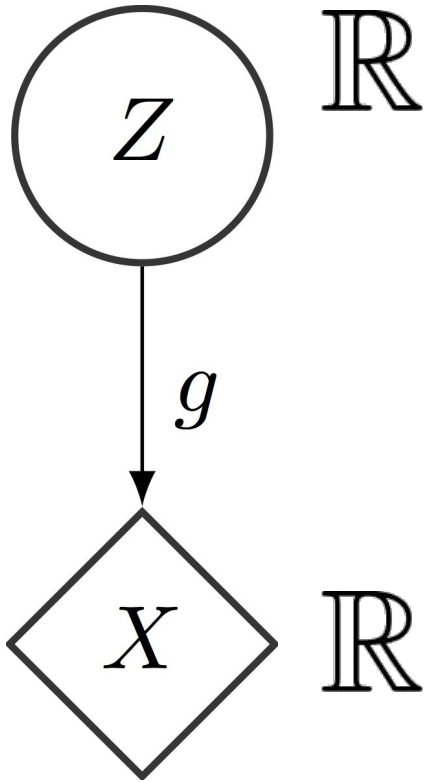
Applications

- Density estimation / conditioning MCMC $f_\theta = g_\theta^{-1}$

$$\log\left(p_X^{(\theta)}(x)\right) = \log\left(p_Z\left(f_\theta(x)\right)\right) + \log\left(\left|\frac{\partial f_\theta}{\partial x}\right|(x)\right)$$



Change of variable formula



$$p_X(x) = p_Z(z) \left| \frac{\partial z}{\partial x} \right|$$

Scalar example

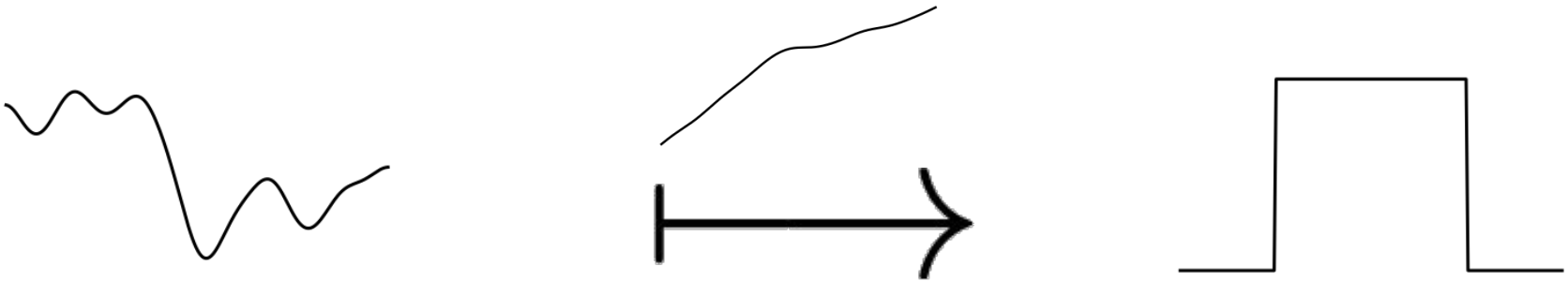
$$\mathcal{N}(x;\mu,\sigma) = \mathcal{N}\left(\frac{x-\mu}{\sigma};0,1\right)\sigma^{-1}$$

(Devroye, 1986)

Scalar example

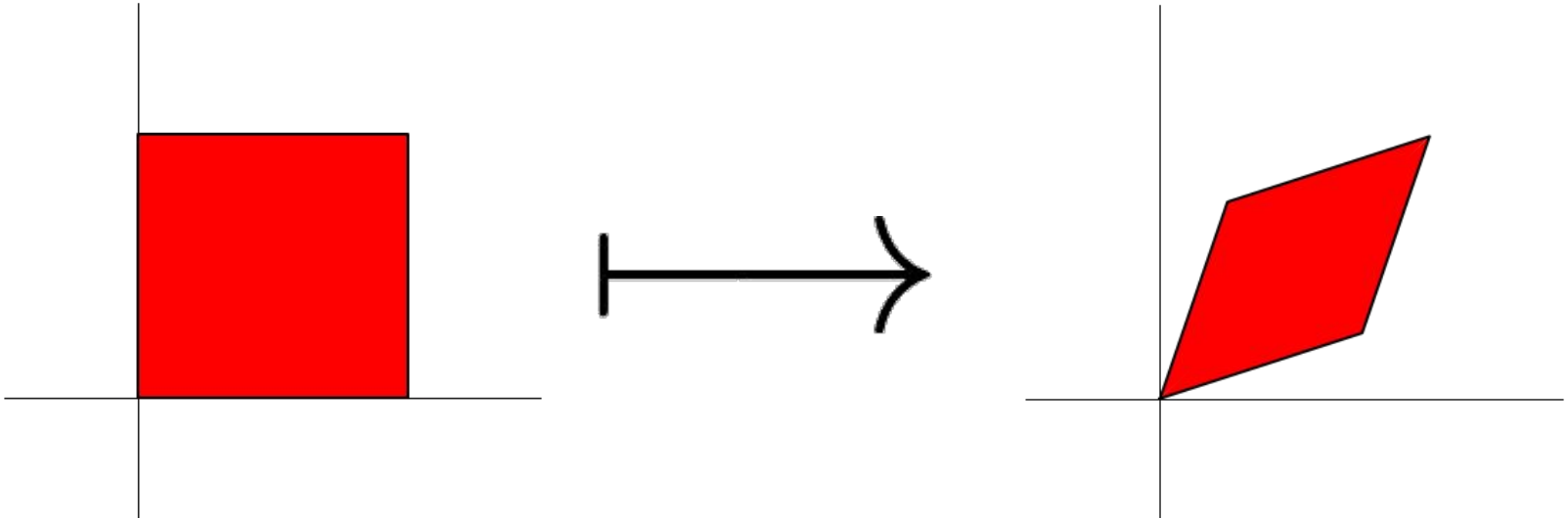
Inverse transform sampling $x \mapsto z = CDF(x)$

$$p_X(x) = \mathcal{U}(z; [0, 1]) \frac{\partial CDF}{\partial x}(x)$$



Multi-dimensional case $\mathbb{R}^d$

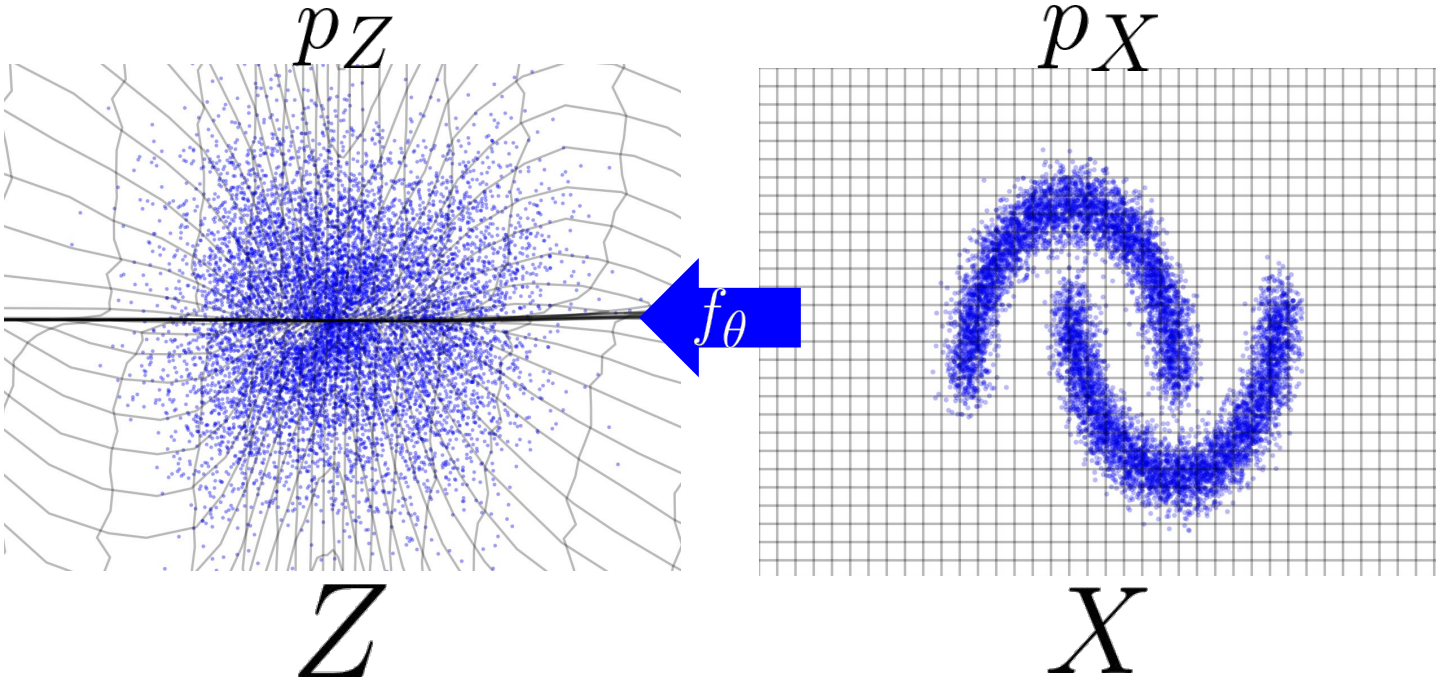
$$p_X(x) = p_Z(z) \left| \det \frac{\partial z}{\partial x} \right| \left(\frac{\partial z}{\partial x} \right) \Bigg|$$



Challenges

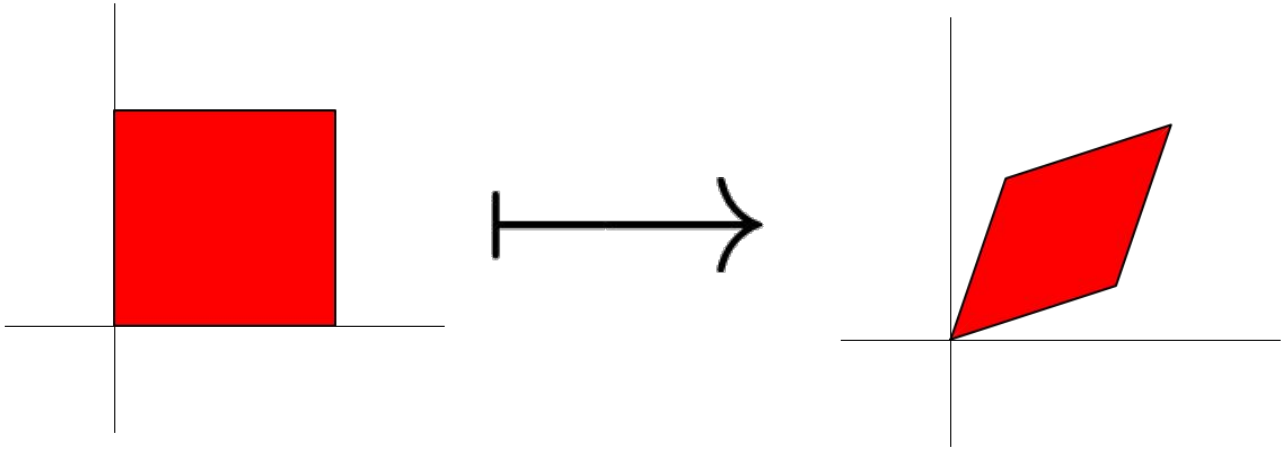
- Jacobian determinant
- Inversion

Study case: density estimation



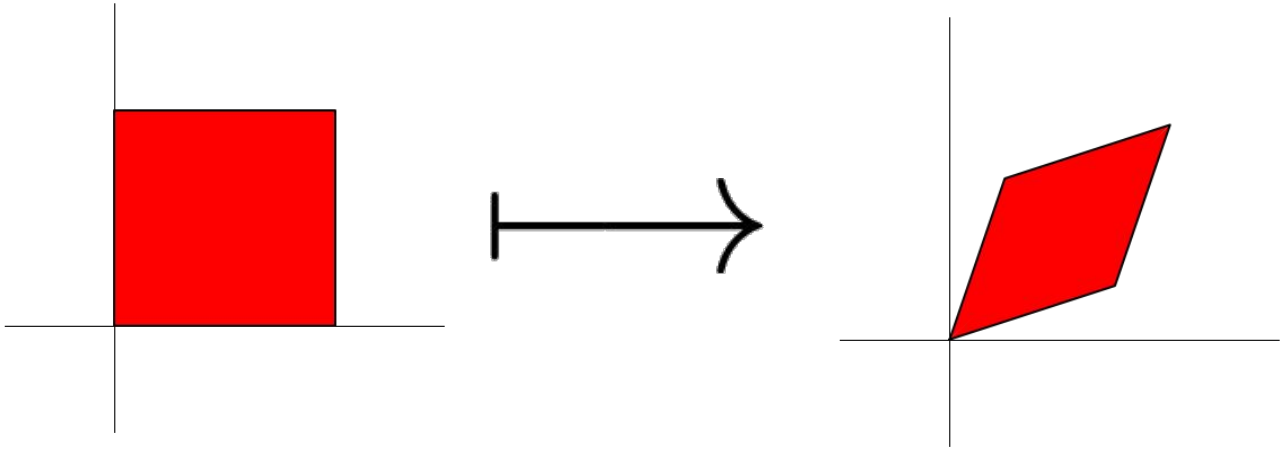
$$\log\left(p_X^{(\theta)}(x)\right) = \log\left(p_Z(f_\theta(x))\right) + \log\left(\left|\frac{\partial f_\theta}{\partial x}\right|(x)\right)$$

Jacobian determinant

- $$\frac{\partial f_\theta}{\partial x} \in \mathbb{M}(d, d)$$
- 

(Chen et al., 2019)

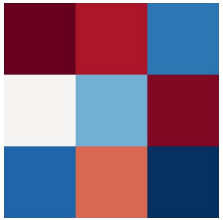
Determinant

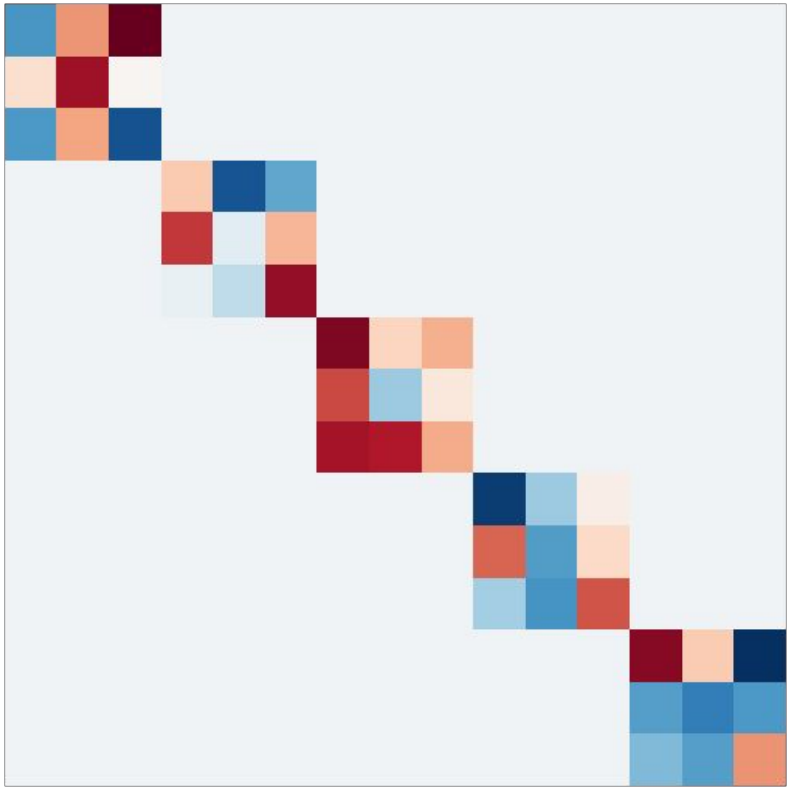


Implementations range from $O(d!)$ to $O(d^3)$ non-parallel

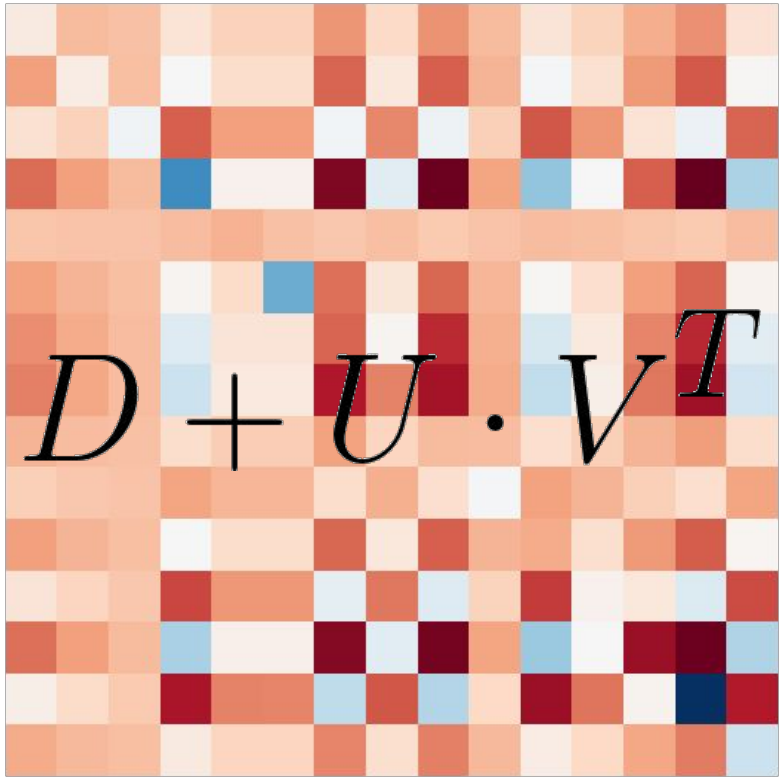
High variance unbiased estimator exists (Hutchinson estimator + Taylor series)

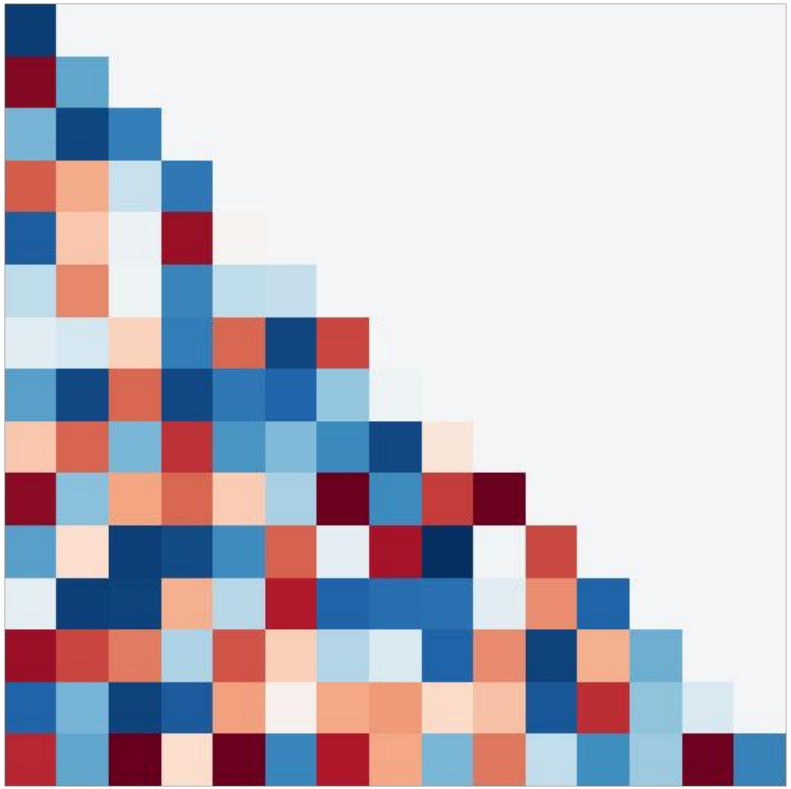
More tractable determinants





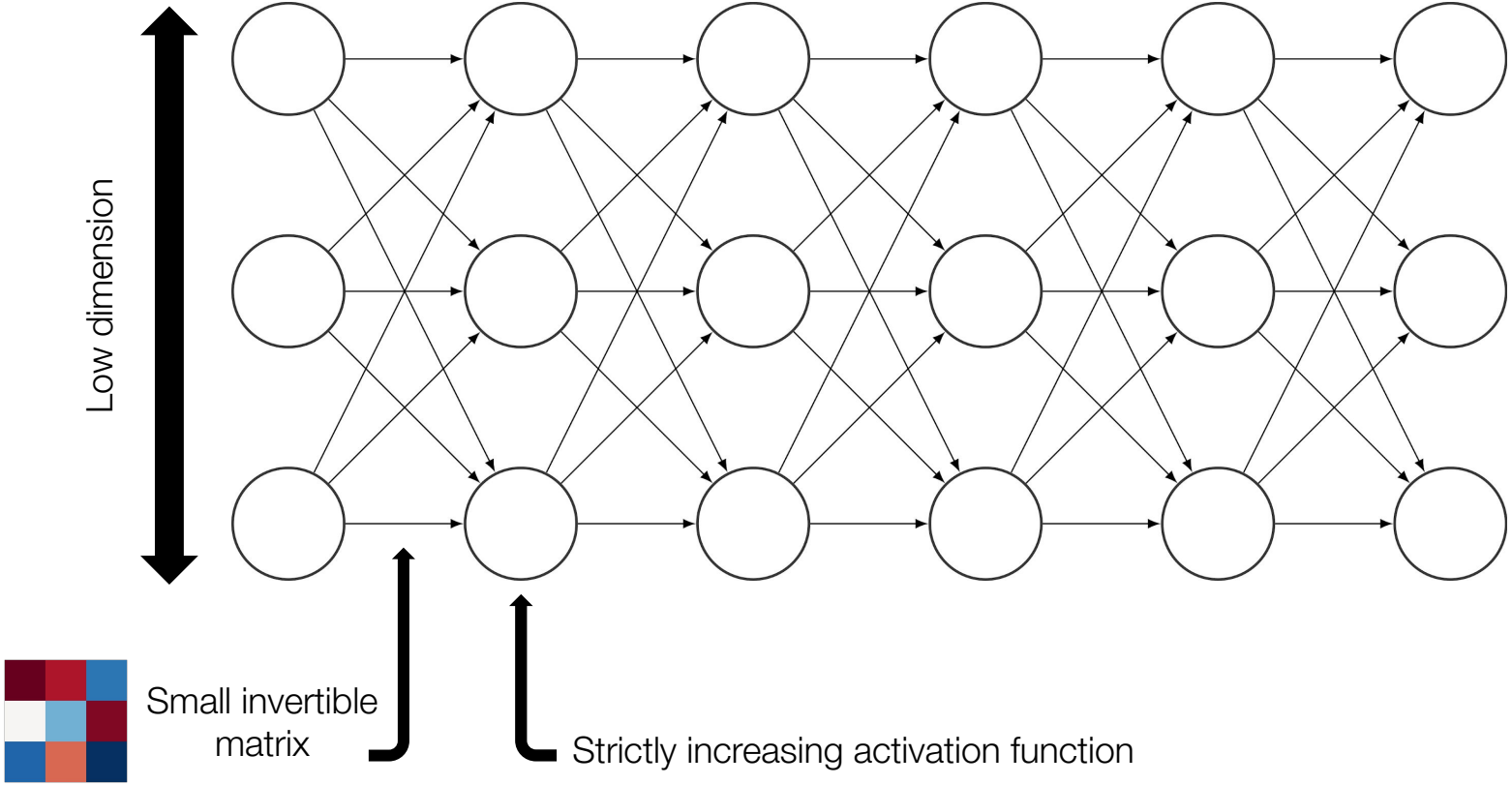
More tractable determinants





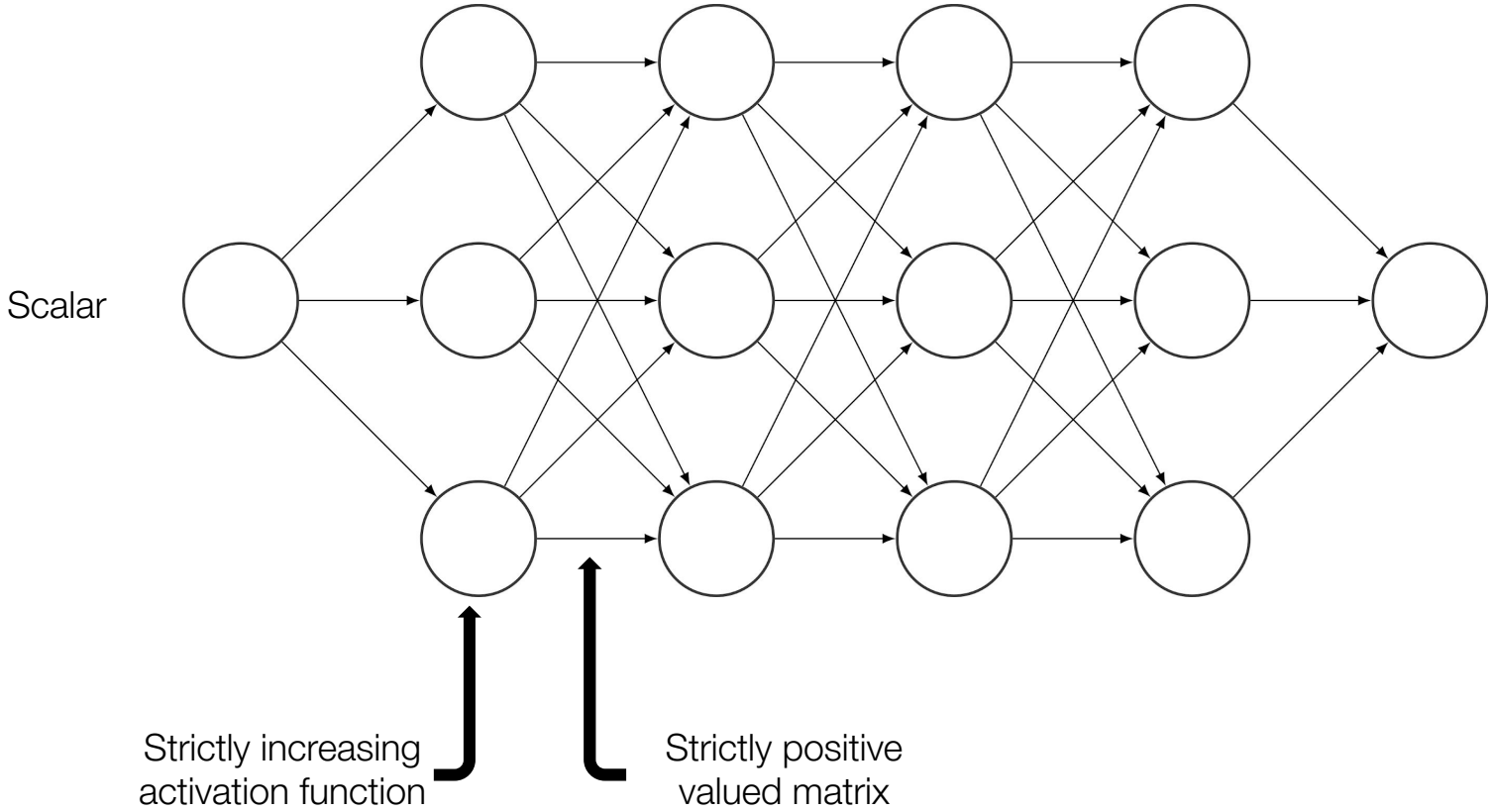
(Baird et al., 2005)

Deep learning with tractable Jacobian determinant



(Huang et al., 2018; De Cao, 2019)

Neural scalar flow

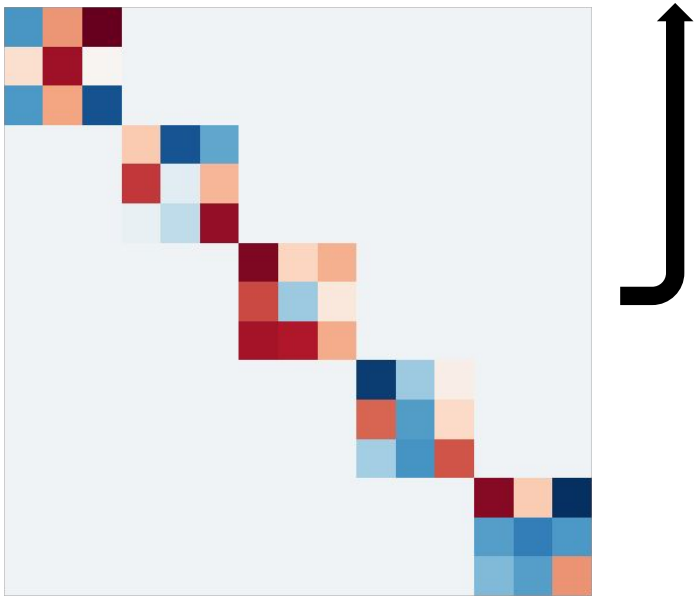


(Hoogeboom et al., 2019; Karami et al., 2019)

Fourier convolution

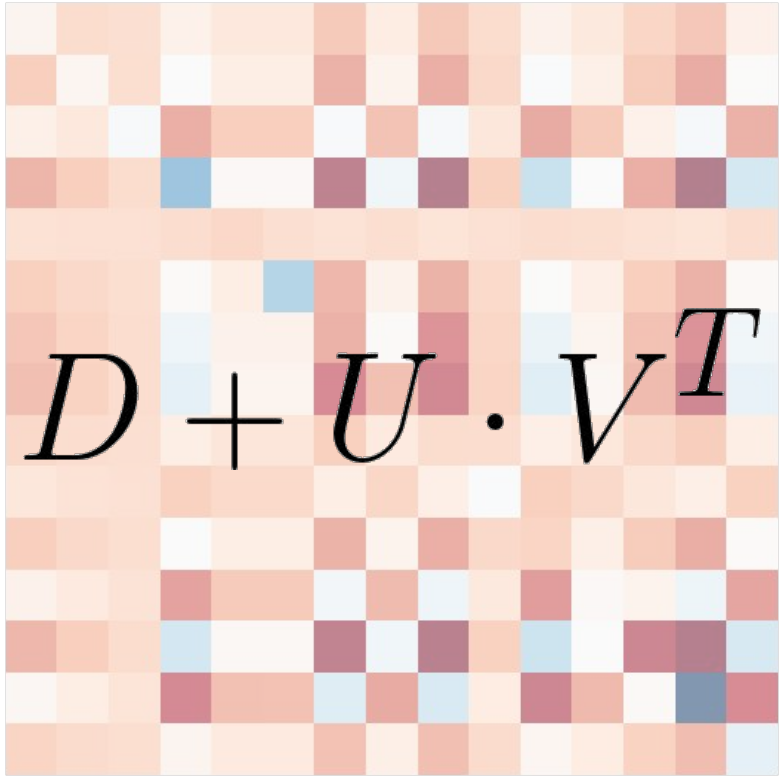
(Periodic) convolution theorem

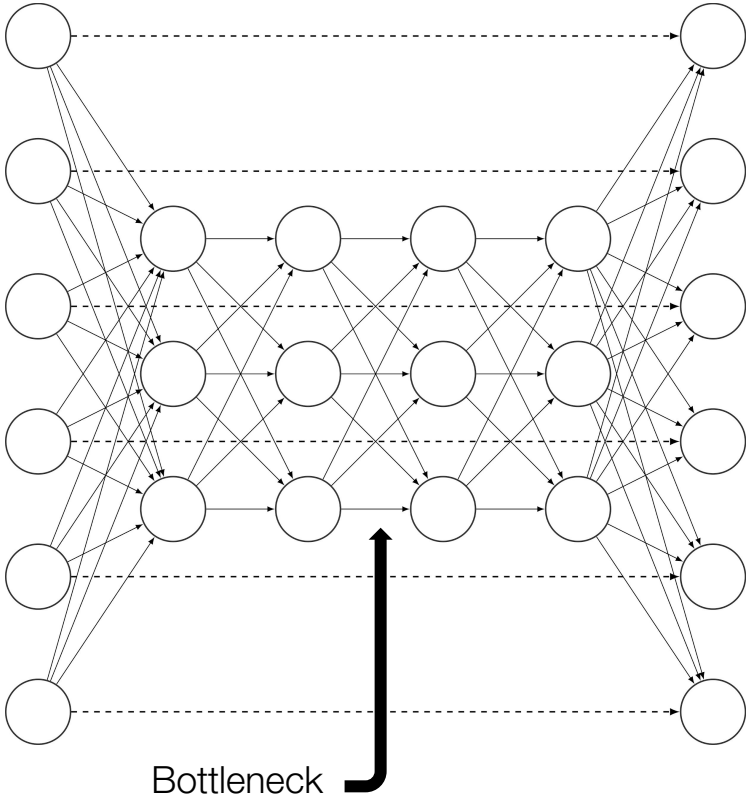
$$\mathcal{F}(x * w) = \mathcal{F}(x) \cdot \mathcal{F}(w)$$



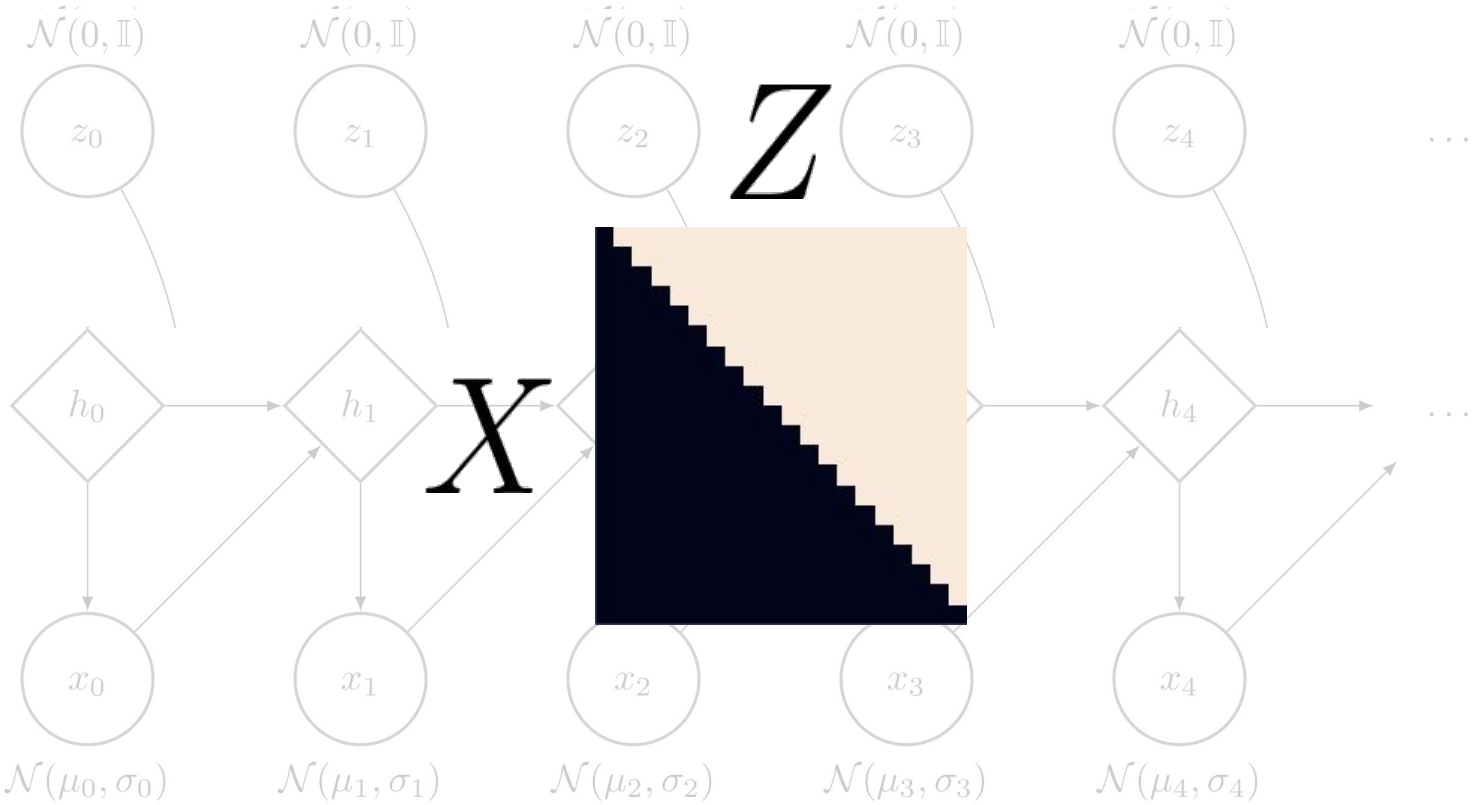
(van den Berg, Hansclever et al., 2018)

Sylvester normalizing flows





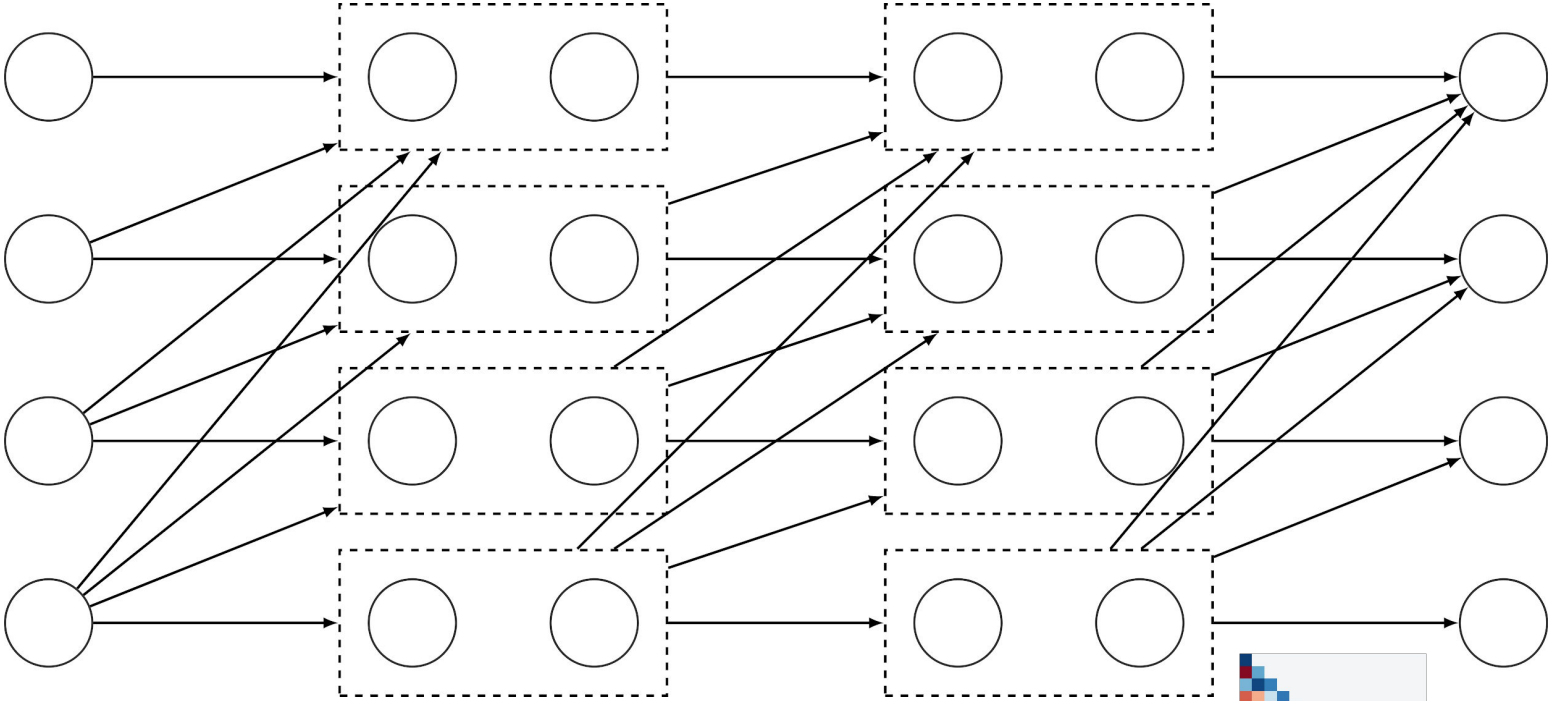
Autoregressive models



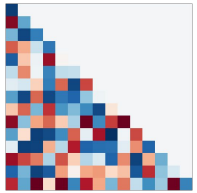
(Deco & Brauer, 1995; Hyvarinen & Pajunen, 1998; Moselhy & Marzouk, 2012)

(Bengio[2], 1999; Larochelle & Murray, 2011; van den Oord et al., 2015; Uria et al., 2016)

Neural autoregressive models



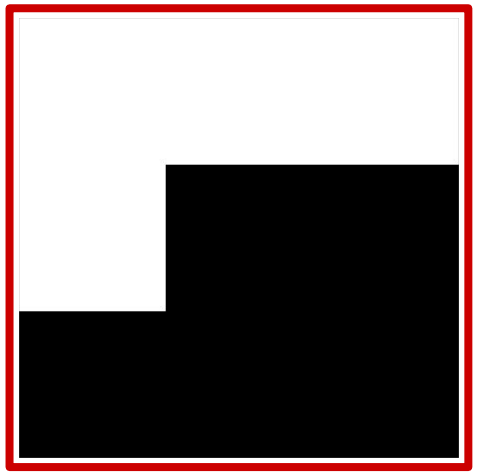
$$f_d(x) = f_d(x_{<d})$$

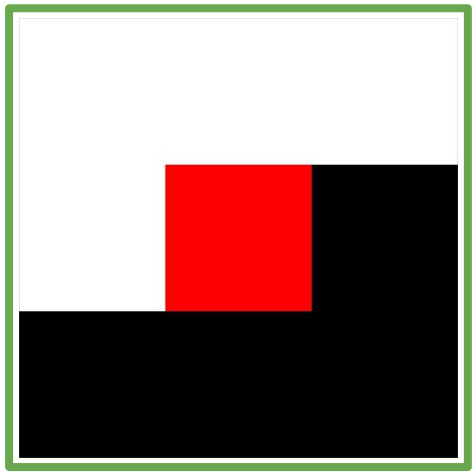


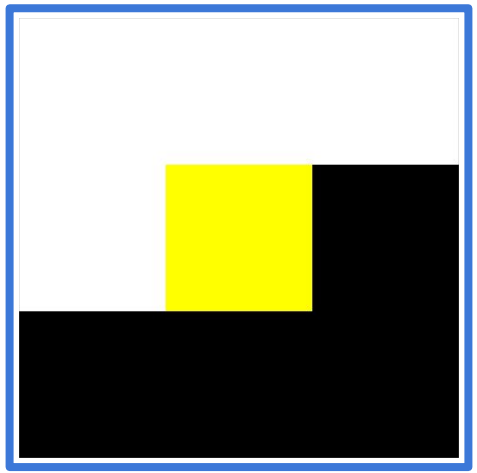
(van den Oord et al., 2016)

Convolutional autoregressive models

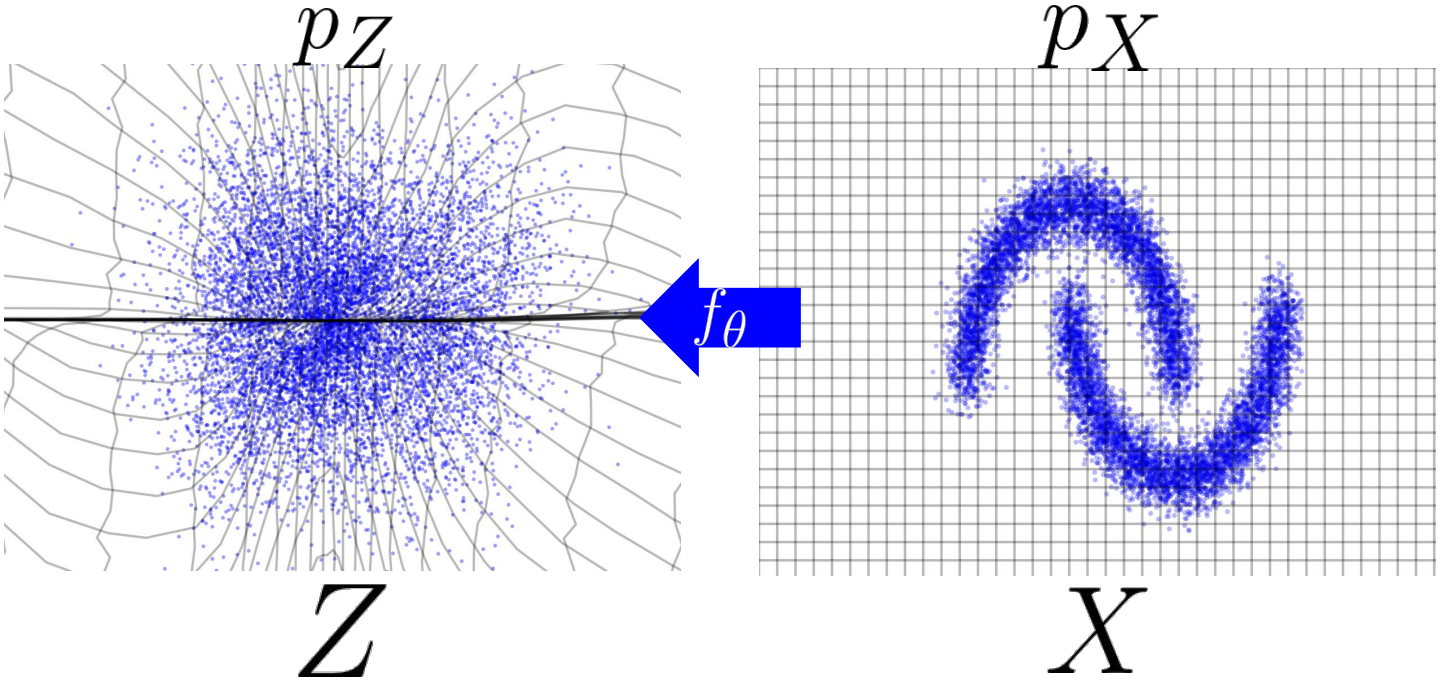
Masked convolutions







Study case: density estimation

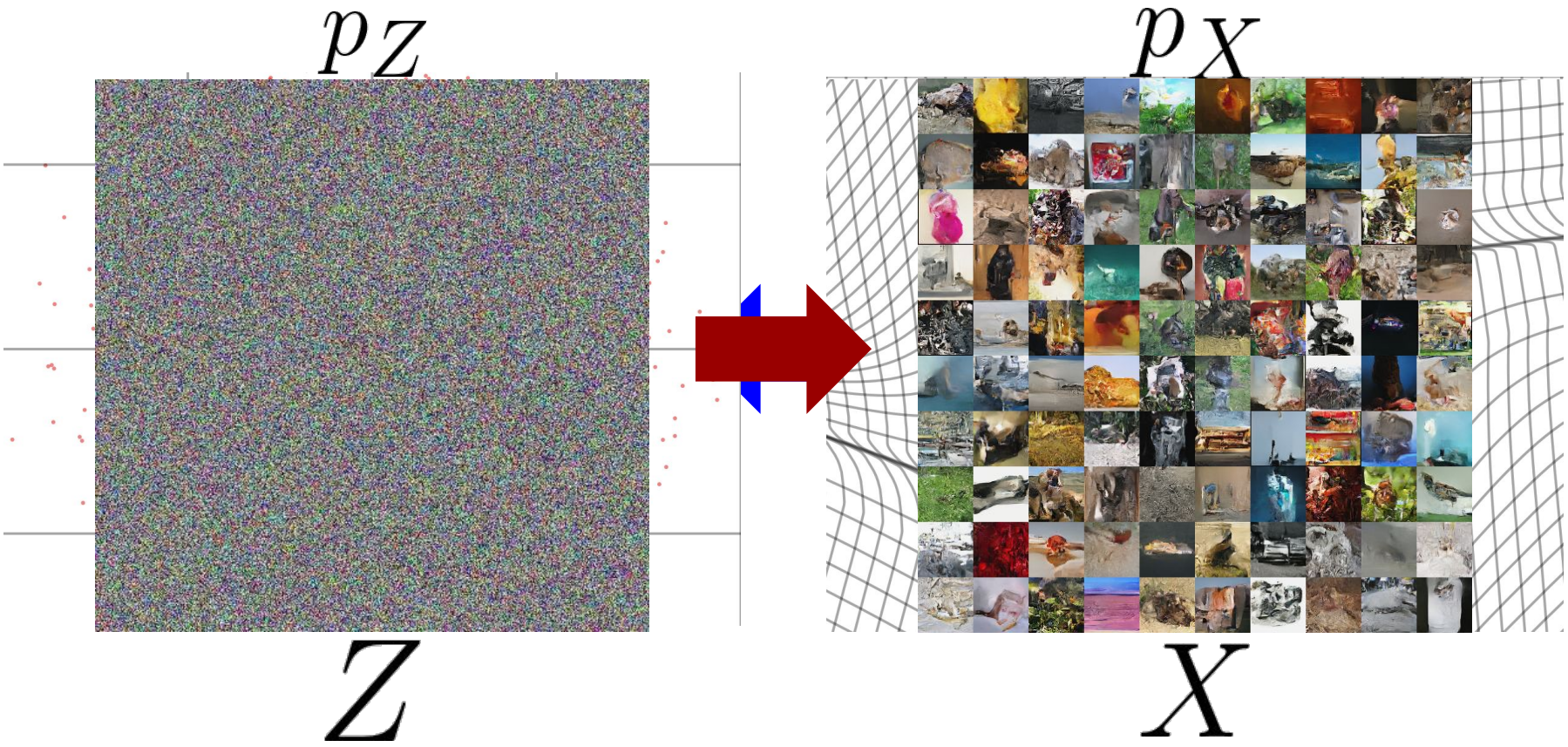


$$\log\left(p_X^{(\theta)}(x)\right) = \log\left(p_Z(f_\theta(x))\right) + \log\left(\left|\frac{\partial f_\theta}{\partial x}\right|(x)\right)$$

Inverting a neural network

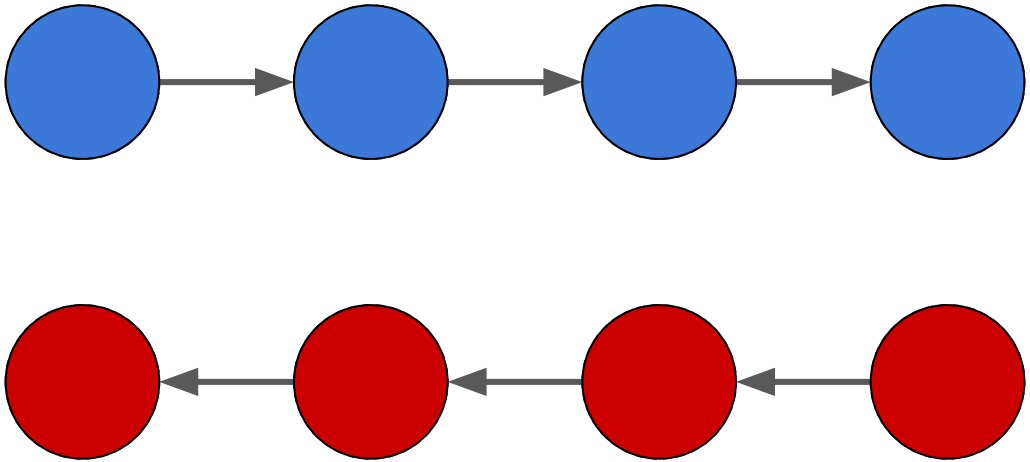
$$f_\theta^{-1}$$

Generation through process reversion



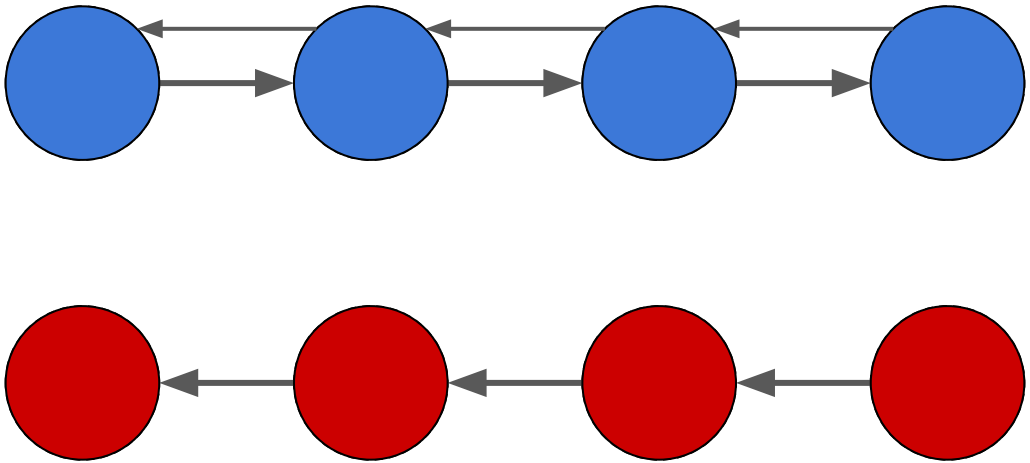
(Gomez, Ren et al., 2017; MacKay et al., 2018)

Reducing backprop memory footprint



(Gomez, Ren et al., 2017; MacKay et al., 2018)

Reducing backprop memory footprint



How to invert a neural net

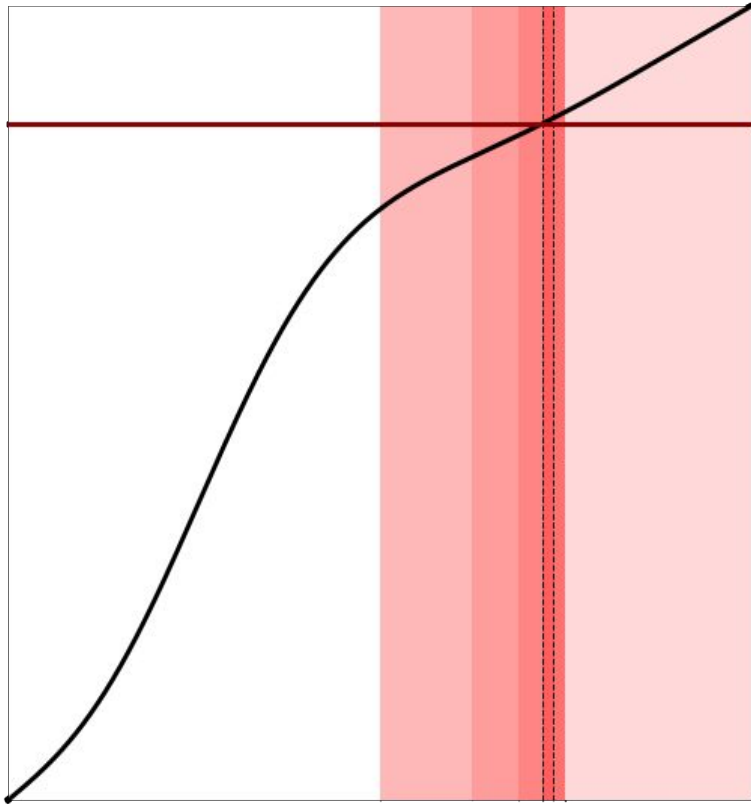
$$f_\theta^{-1} = \ ?$$

Iterative inversion

- Bisection / binary search
- Root finding algorithm (Newton Raphson)
- Fixed point iteration

(Ho, Chen et al., 2019)

Bisection



(Song et al., 2019)

Root finding algorithm

Newton-Raphson

$$x^{(t+1)} = x^{(t)} - \alpha \left(\frac{\partial f}{\partial x} \right)^{-1} \left(f(x^{(t)}) - y \right)$$

Local convergence

(Behrmann et al., 2019)

Residual flow

$$x \mapsto x + f(x) = y$$

$$\|f(x^{(1)}) - f(x^{(2)})\| \leq c\|x^{(1)} - x^{(2)}\|$$

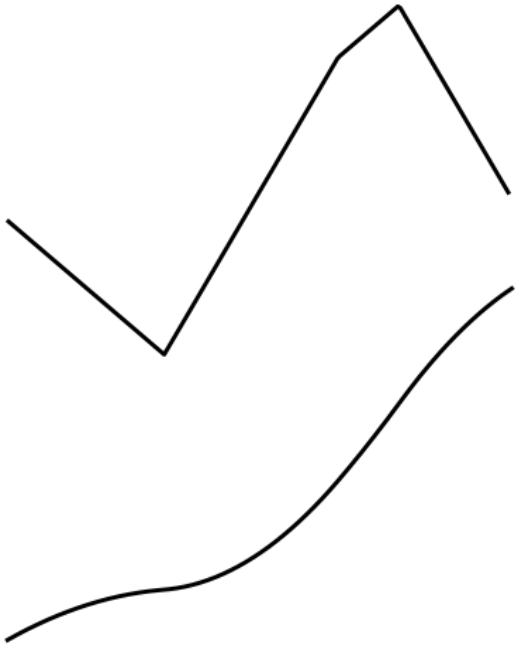
$$x^{(t+1)} = y - f(x^{(t)})$$

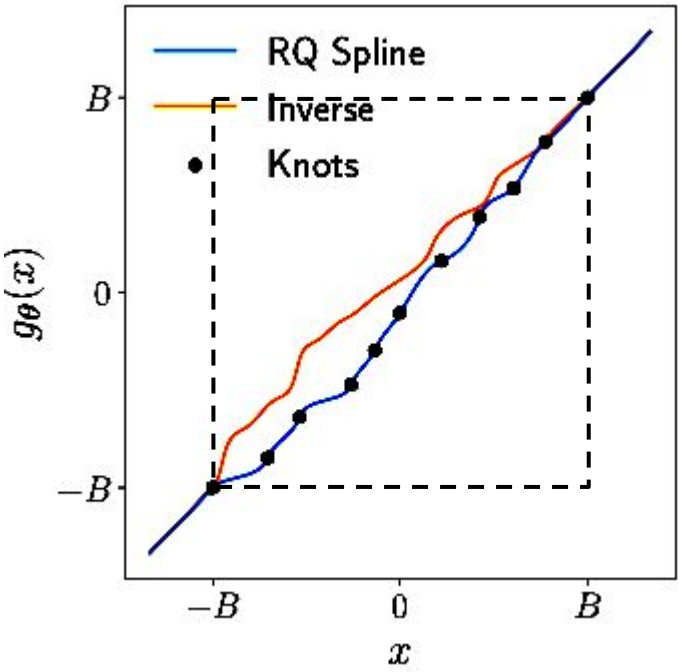
Global convergence

(Müller et al., 2019; Durkan, Bekasov et al., 2019)

Closed form inverse: scalar case

Invertible piecewise functions





Autoregressive case

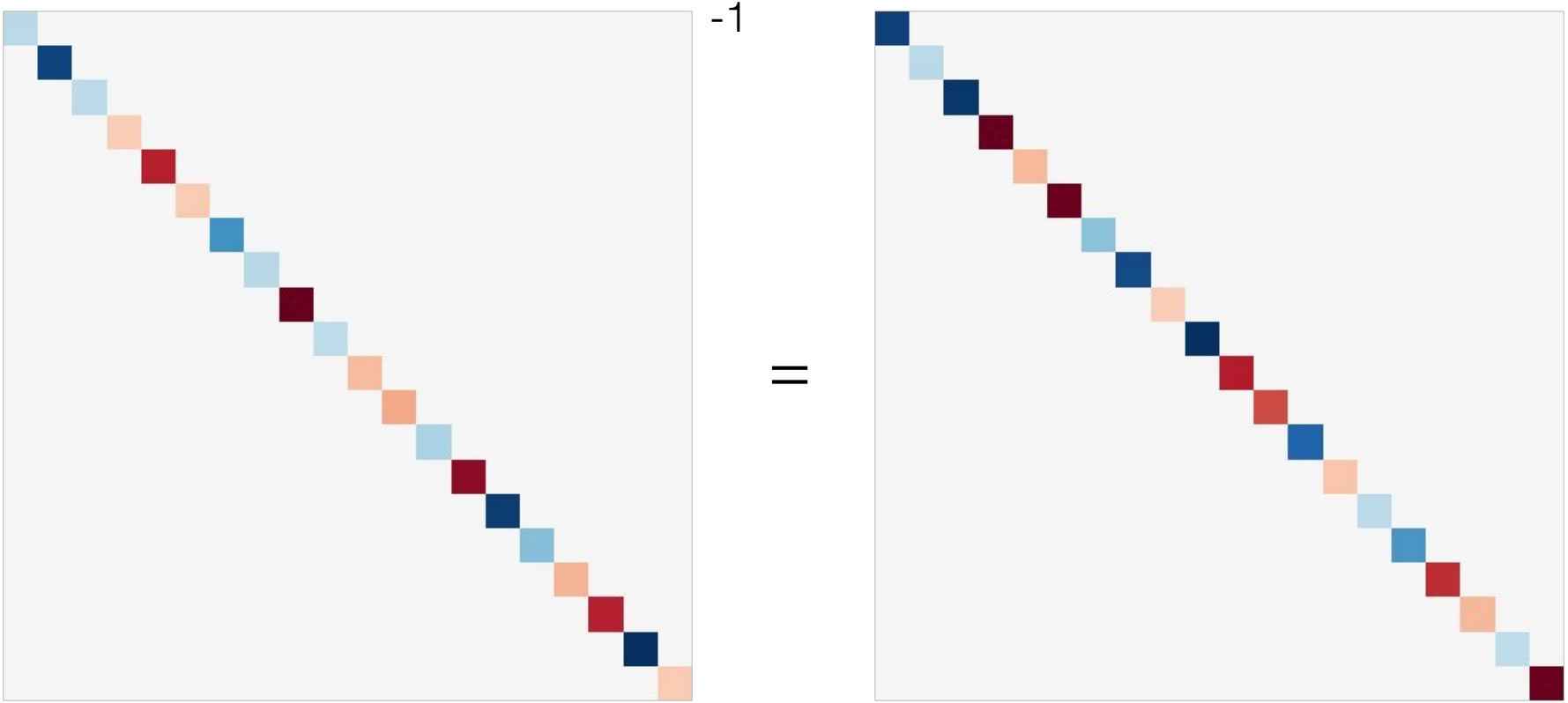
Forward substitution

$$z_d = f_d(x_d; x_{<d})$$

$$x_d = f_d^{-1}(z_d; x_{<d})$$

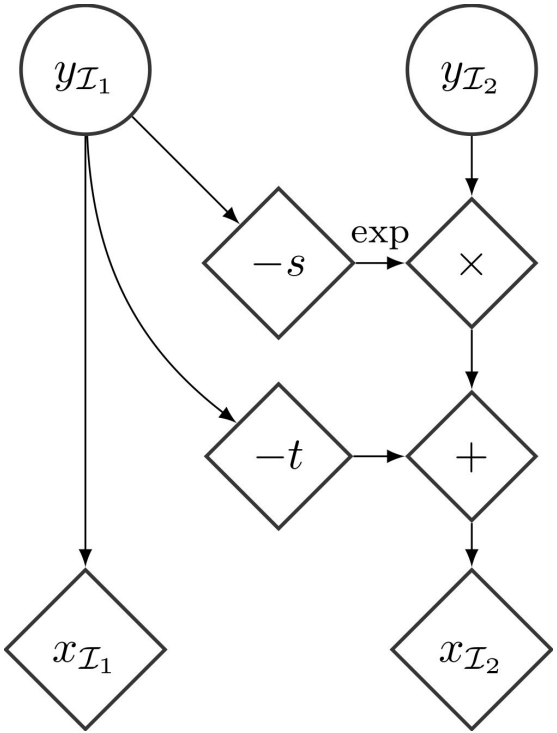
Non parallel

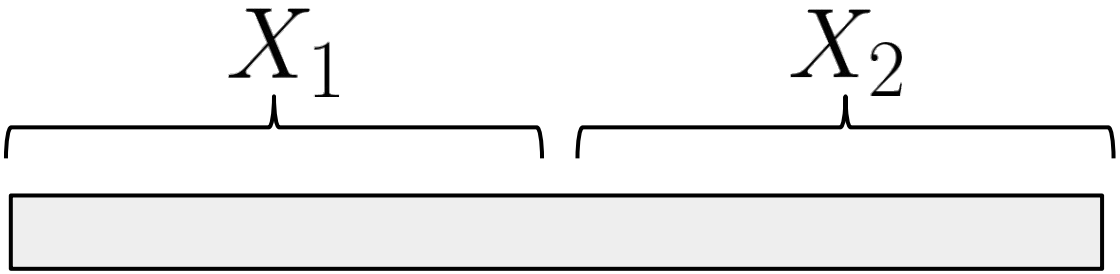
Easier inverse



(Dinh et al., 2014; Dinh et al., 2016)

Coupling layer

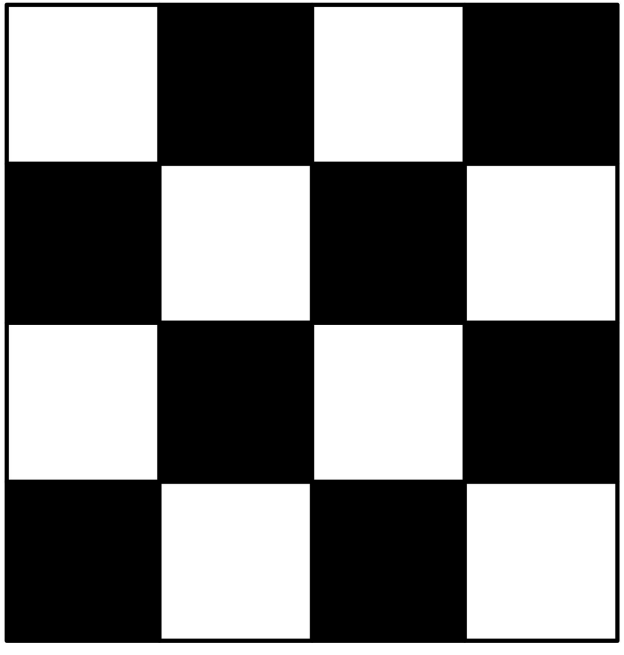


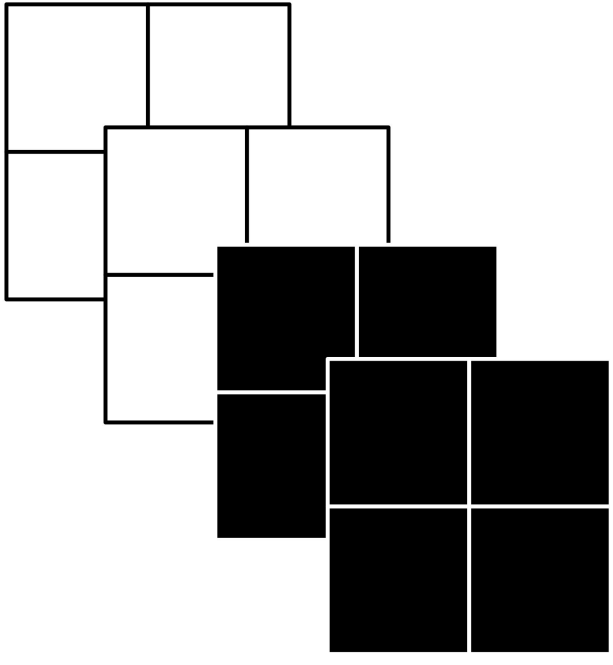


$$X_1 = Y_1$$

$$X_2 = Y_2 \odot \exp\left(s(Y_1)\right) + t(Y_1)$$

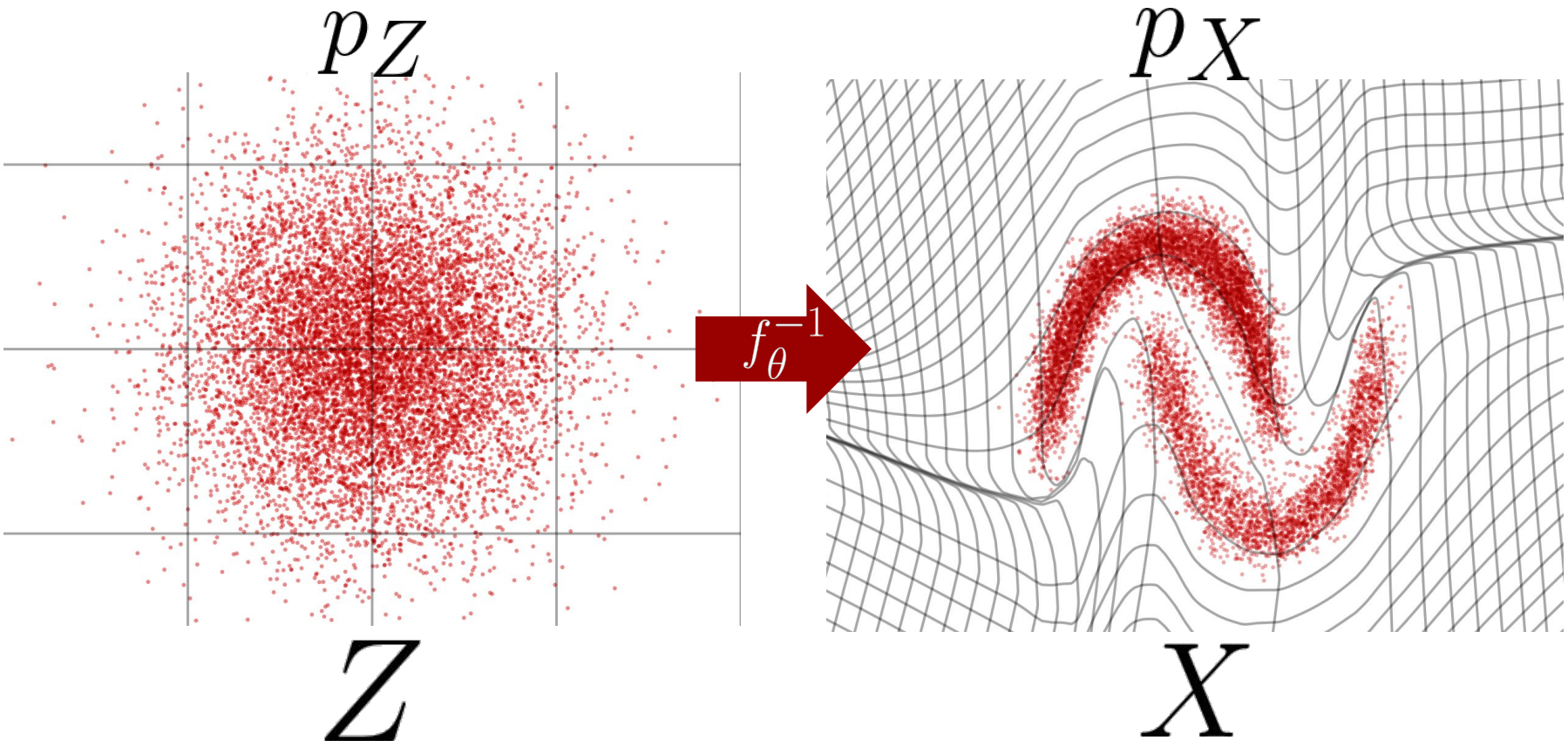
Convolution compatible masking

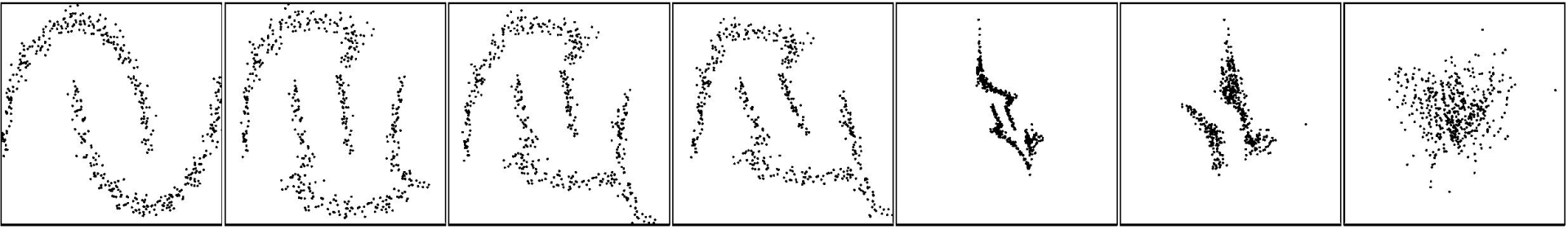




$$Y = (X + (1 - b) \odot t(b \odot X)) \odot \exp\left((1 - b) \odot s(b \odot X)\right)$$

Generation through process reversion





Composing flows

$$f_3 \circ f_2 \circ f_1$$

Composing flows

- Inversion and sampling

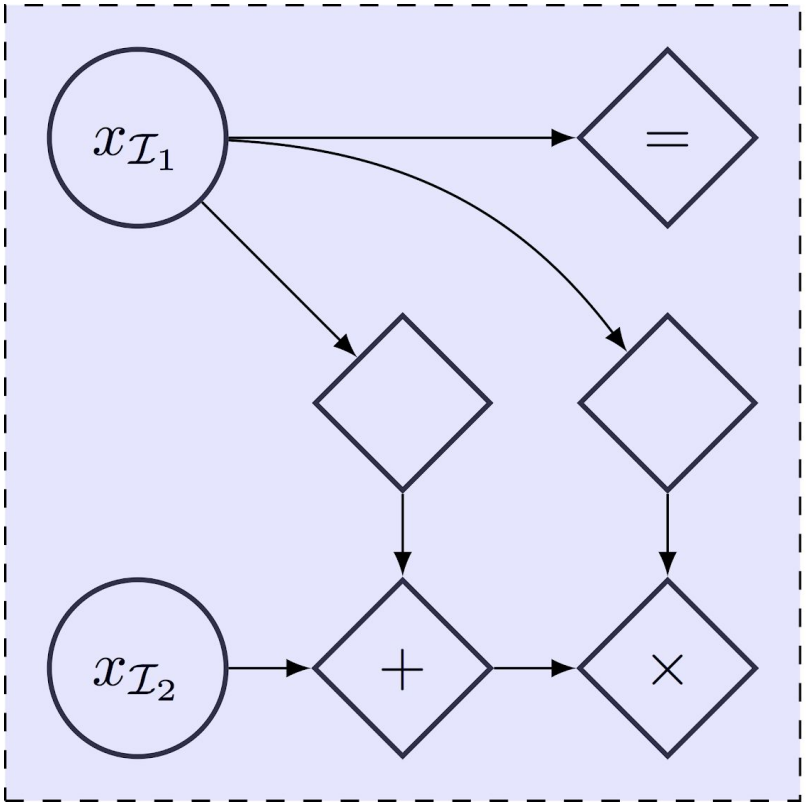
$$(f_2 \circ f_1)^{-1} = f_1^{-1} \circ f_2^{-1}$$

- Determinant and inference

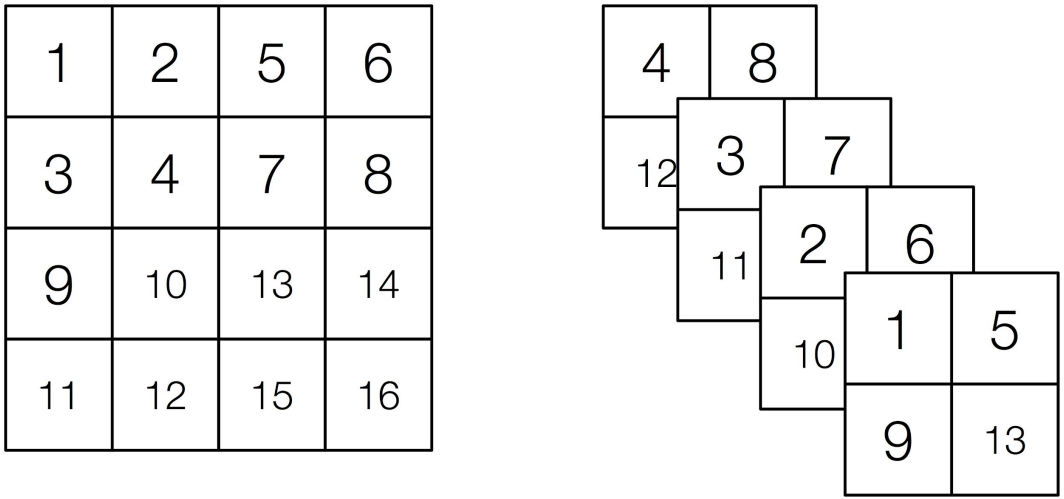
$$\nabla(f_2 \circ f_1)(x) = \nabla f_2\big(f_1(x)\big)\nabla f_1(x)$$

$$\det(A \cdot B) = \det(A) \cdot \det(B)$$

Combining coupling layers



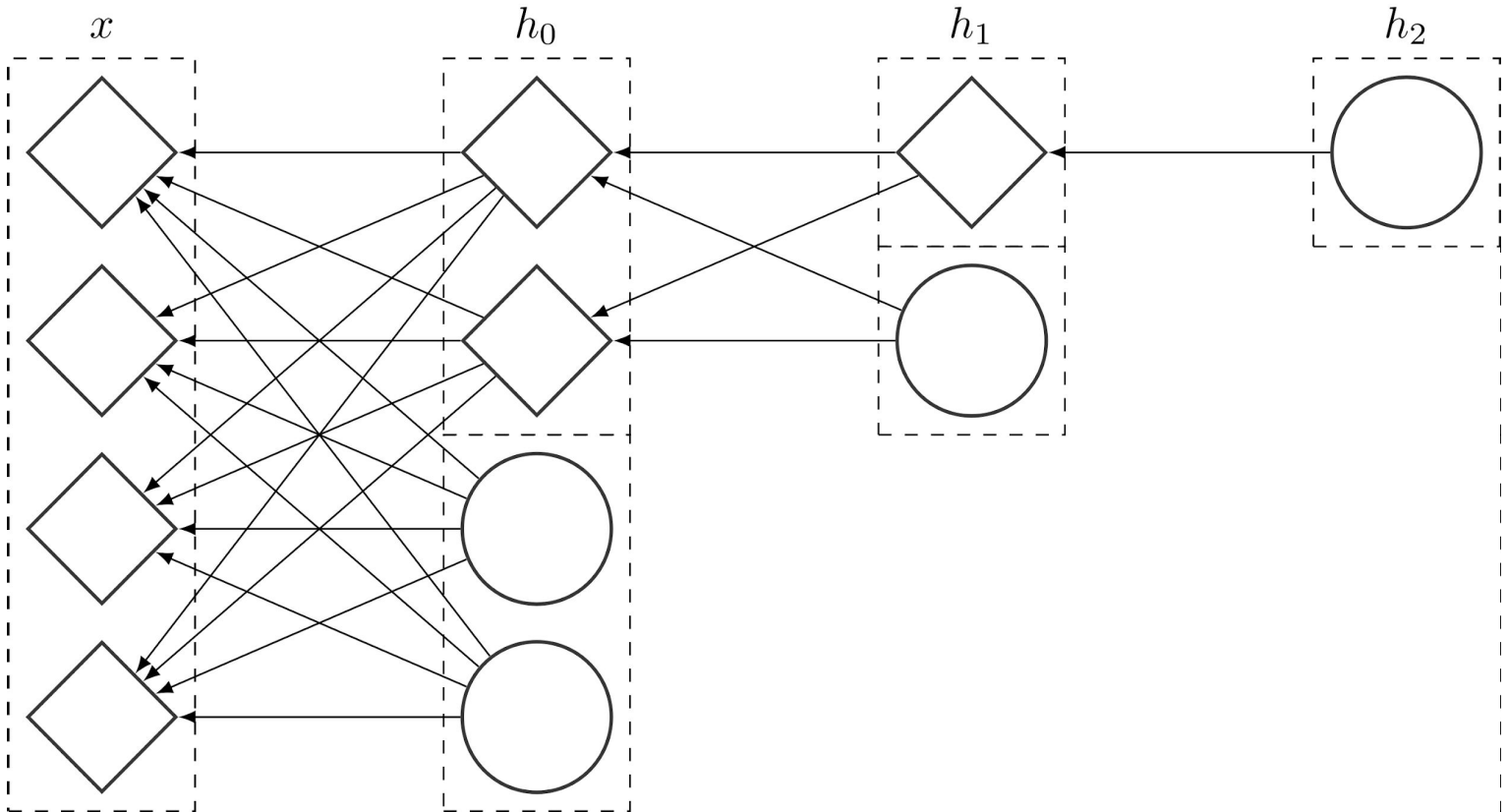
Multi-scale architecture: downsampling



Relate to:

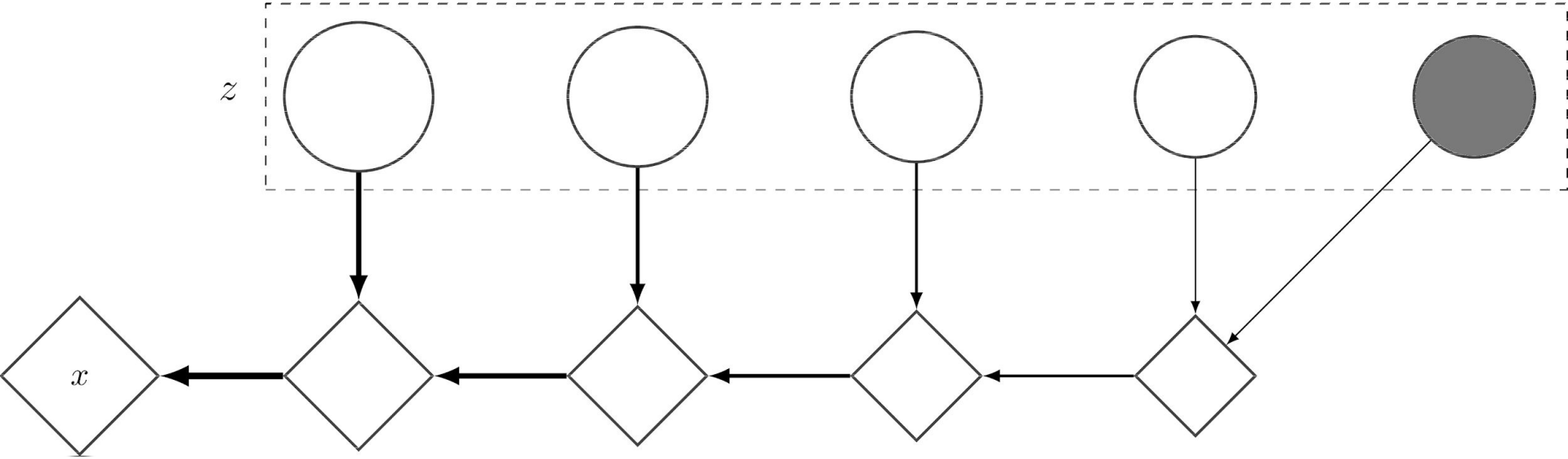
- Strided convolution
- Dilated convolution

Managing the bijectivity constraint

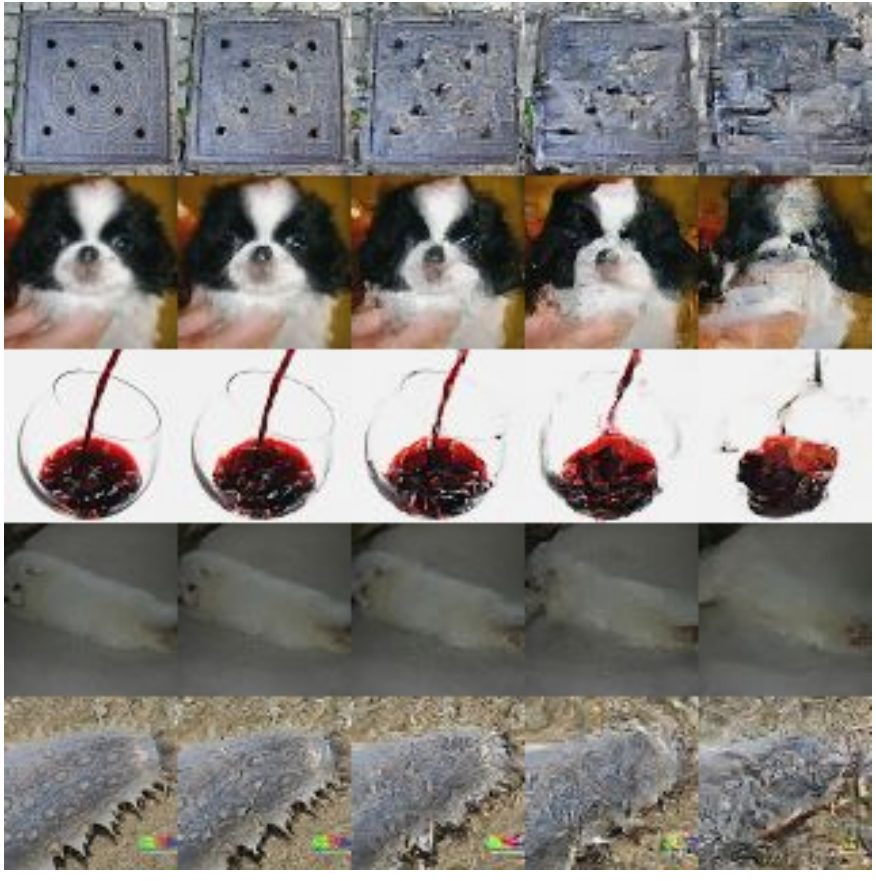


(Gregor et al., 2016)

Multi-layer architecture



Conceptual compression



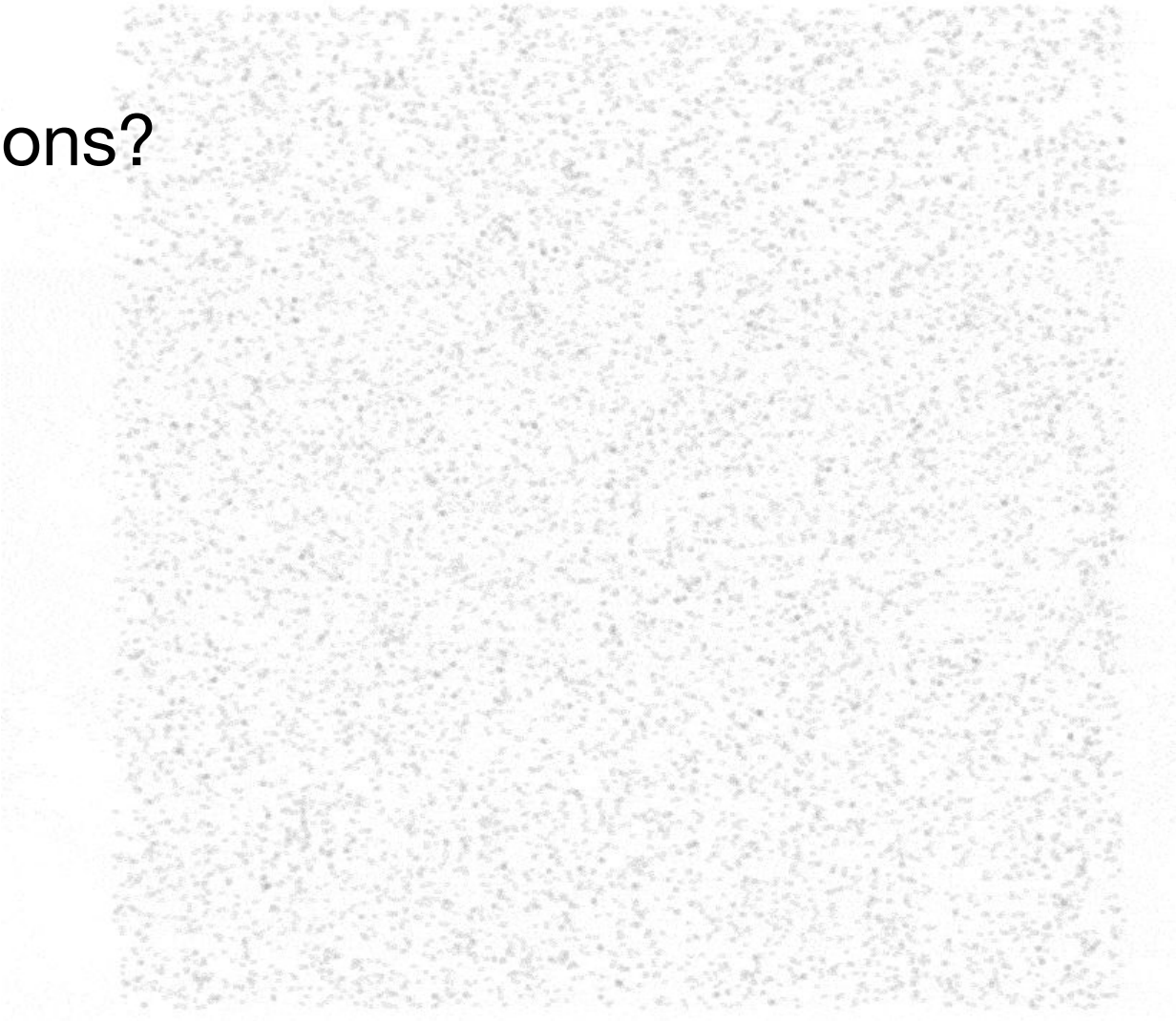
(Chen & Rubanova & Bettencourt et al., 2018; Grathwohl & Chen et al., 2018)

Research in progress

- Continuous time flows
- Discrete value flows
- Adaptive sparsity structure
- Non-invertible flows

(Tran et al., 2019; Hoogeboom et al., 2019; Dinh et al., 2019)

Questions?



References

- Kingma, Durk P., and Prafulla Dhariwal. "Glow: Generative flow with invertible 1x1 convolutions." Advances in Neural Information Processing Systems. 2018.
- Hyvärinen, Aapo, and Erkki Oja. "Independent component analysis: algorithms and applications." Neural networks 13.4-5 (2000): 411-430.
- Rippel, Oren, and Ryan Prescott Adams. "High-dimensional probability estimation with deep density models." arXiv preprint arXiv:1302.5125 (2013).
- Dinh, Laurent, David Krueger, and Yoshua Bengio. "Nice: Non-linear independent components estimation." arXiv preprint arXiv:1410.8516 (2014).
- Hoffman, Matthew, Pavel Sountsov, Joshua V. Dillon, Ian Langmore, Dustin Tran, and Srinivas Vasudevan. "NeuTra-lizing Bad Geometry in Hamiltonian Monte Carlo Using Neural Transport." arXiv preprint arXiv:1903.03704 (2019).

References

- Hoffman, Matthew D., David M. Blei, Chong Wang, and John Paisley. "Stochastic variational inference." The Journal of Machine Learning Research 14, no. 1 (2013): 1303-1347.
- Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
- Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate inference in deep generative models." arXiv preprint arXiv:1401.4082 (2014).
- Devroye, Luc. "Nonuniform random variate generation." Handbooks in operations research and management science 13 (2006): 83-121.
- Chen, Ricky TQ, Jens Behrmann, David Duvenaud, and Jörn-Henrik Jacobsen. "Residual Flows for Invertible Generative Modeling." arXiv preprint arXiv:1906.02735 (2019).

References

- Baird, Leemon, David Smalenberger, and Shawn Ingkiriwang. "One-step neural network inversion with PDF learning and emulation." Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.. Vol. 2. IEEE, 2005.
- Huang, Chin-Wei, David Krueger, Alexandre Lacoste, and Aaron Courville. "Neural autoregressive flows." arXiv preprint arXiv:1804.00779 (2018).
- De Cao, Nicola, Ivan Titov, and Wilker Aziz. "Block neural autoregressive flow." arXiv preprint arXiv:1904.04676 (2019).
- Hoogeboom, Emiel, Rianne van den Berg, and Max Welling. "Emerging convolutions for generative normalizing flows." arXiv preprint arXiv:1901.11137 (2019).
- Karami, Mahdi, Sohl-Dickstein, Jascha, Dinh, Laurent, Duckworth, Daniel, and Schuurmans, Dale “Symmetric Convolutional Flow”.

References

- Berg, Rianne van den, Leonard Hasenclever, Jakub M. Tomczak, and Max Welling. "Sylvester normalizing flows for variational inference." arXiv preprint arXiv:1803.05649 (2018).
- Deco, Gustavo, and Wilfried Brauer. "Nonlinear higher-order statistical decorrelation by volume-conserving neural architectures." Neural Networks 8.4 (1995): 525-535.
- Hyvarinen, A., and Petteri Pajunen. "On existence and uniqueness of solutions in nonlinear independent component analysis." 1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227). Vol. 2. IEEE, 1998.
- El Moselhy, Tarek A., and Youssef M. Marzouk. "Bayesian inference with optimal maps." Journal of Computational Physics 231.23 (2012): 7815-7850.

References

- Bengio, Yoshua, and Samy Bengio. "Modeling high-dimensional discrete data with multi-layer neural networks." Advances in Neural Information Processing Systems. 2000.
- Larochelle, Hugo, and Iain Murray. "The neural autoregressive distribution estimator." Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. 2011.
- Oord, Aaron van den, Nal Kalchbrenner, and Koray Kavukcuoglu. "Pixel recurrent neural networks." arXiv preprint arXiv:1601.06759 (2016).
- Uria, Benigno, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. "Neural autoregressive distribution estimation." The Journal of Machine Learning Research 17, no. 1 (2016): 7184-7220.
- Van den Oord, Aaron, et al. "Conditional image generation with pixelcnn decoders." Advances in neural information processing systems. 2016.

References

- Gomez, Aidan N., Mengye Ren, Raquel Urtasun, and Roger B. Grosse. "The reversible residual network: Backpropagation without storing activations." In Advances in neural information processing systems, pp. 2214-2224. 2017.
- MacKay, Matthew, et al. "Reversible recurrent neural networks." Advances in Neural Information Processing Systems. 2018.
- Ho, Jonathan, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. "Flow++: Improving flow-based generative models with variational dequantization and architecture design." arXiv preprint arXiv:1902.00275 (2019).
- Song, Yang, Chenlin Meng, and Stefano Ermon. "MintNet: Building Invertible Neural Networks with Masked Convolutions." arXiv preprint arXiv:1907.07945 (2019).
- Behrmann, Jens, David Duvenaud, and Jörn-Henrik Jacobsen. "Invertible residual networks." arXiv preprint arXiv:1811.00995 (2018).

References

- Müller, Thomas, Brian McWilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. "Neural importance sampling." arXiv preprint arXiv:1808.03856 (2018).
- Durkan, Conor, Artur Bekasov, Iain Murray, and George Papamakarios. "Neural Spline Flows." arXiv preprint arXiv:1906.04032 (2019).
- Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using real nvp." arXiv preprint arXiv:1605.08803 (2016).
- Gregor, Karol, Frederic Besse, Danilo Jimenez Rezende, Ivo Danihelka, and Daan Wierstra. "Towards conceptual compression." In Advances In Neural Information Processing Systems, pp. 3549-3557. 2016.

References

- Chen, Tian Qi, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. "Neural ordinary differential equations." In Advances in neural information processing systems, pp. 6571-6583. 2018.
- Grathwohl, Will, Ricky TQ Chen, Jesse Betterncourt, Ilya Sutskever, and David Duvenaud. "Ffjord: Free-form continuous dynamics for scalable reversible generative models." arXiv preprint arXiv:1810.01367 (2018).
- Tran, Dustin, Keyon Vafa, Kumar Krishna Agrawal, Laurent Dinh, and Ben Poole. "Discrete Flows: Invertible Generative Models of Discrete Data." arXiv preprint arXiv:1905.10347 (2019).
- Hoogeboom, Emiel, Jorn WT Peters, Rianne van den Berg, and Max Welling. "Integer Discrete Flows and Lossless Compression." arXiv preprint arXiv:1905.07376 (2019).
- Dinh, Laurent, Jascha Sohl-Dickstein, Razvan Pascanu, and Hugo Larochelle. "A RAD approach to deep mixture models." arXiv preprint arXiv:1903.07714 (2019).

References

- Noé, Frank, Simon Olsson, Jonas Köhler, and Hao Wu. "Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning." Science 365, no. 6457 (2019): eaaw1147.
- Cranmer, Kyle, Siavash Golkar, and Duccio Pappadopulo. "Inferring the quantum density matrix with machine learning." arXiv preprint arXiv:1904.05903 (2019).
- Ballé, Johannes, Valero Laparra, and Eero P. Simoncelli. "Density modeling of images using a generalized normalization transformation." arXiv preprint arXiv:1511.06281 (2015).