# Variational Inference with Normalising Flows
# 1st Presentation

Yang Chen, Qihang Tian

TU Berlin

December 5, 2018

# Overview

# Variational Inference

# What is Variational Inference

## Main Idea:

The intractable posterior distributions $p(\mathbf{z}|\mathbf{x})$ are approximated by a class of well-known, simple probability distribution families $q(\mathbf{z})$, over which we search for the best approximation to the true posterior.

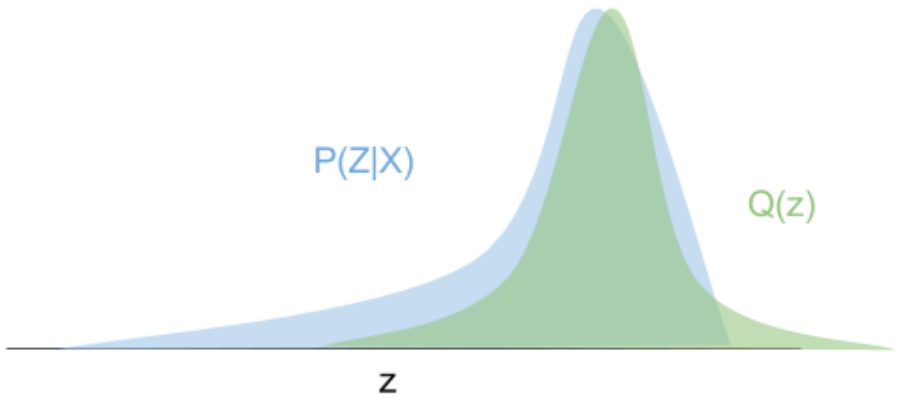



Figure: Approximating true posterior $P(Z|X)$ by normal distribution $Q(Z)$

# Two Commonly Experienced Problems

There are a lots of evidences that describe the detrimental effect of limited posterior approximations, since the class of approximations used is often limited, e.g. mean-field approximations.

Turner & Sahani(2011) summarized two commonly experienced problems:

- Under-estimation of the variance of the posterior distribution
- Limited capacity of the posterior approximation can also result in biases in the MAP estimates of any model parameters

# An Example of Under-estimation of Variance

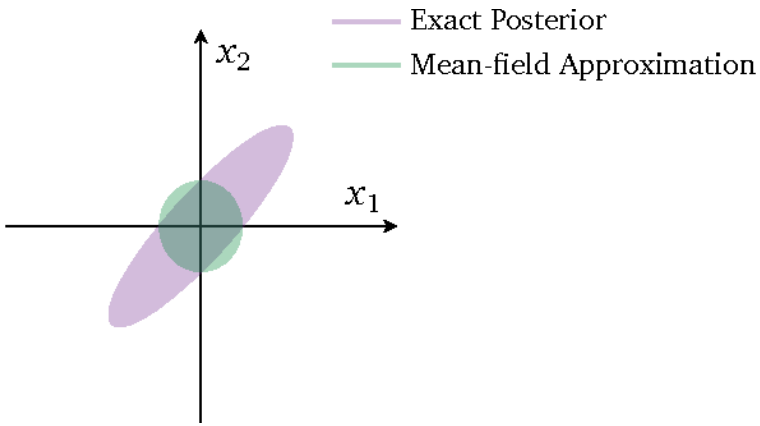



Figure: Visualizing the mean-field approximation to a 2d Gaussian posterior

The ellipses show the effect of mean-field factorization. Since it cannot capture correlation between those latent variables, the marginal variances of the approximation under-represent those of the true density.

# Normalizing Flows

# Normalizing Flows

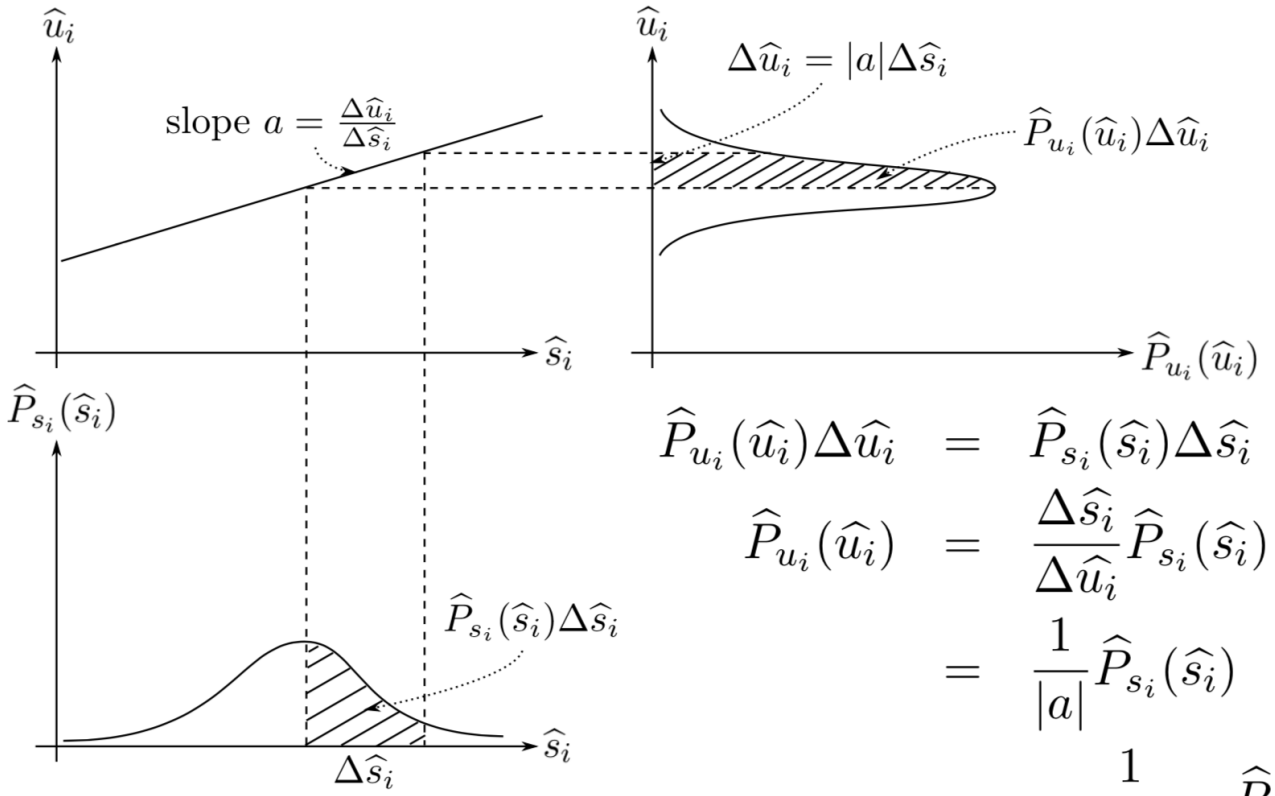
An ideal family of variational distributions $q_\phi(\mathbf{z}|\mathbf{x})$ is the one that highly flexible, preferably flexible enough to obtain the true posterior in an asymptotic regime.

One path towards this goal is based on the principle of normalizing flows (Tabak & Turner, 2013; Tabak & Vanden-Eijnden, 2010):

**Definition:**

A normalizing flow describes the transformation from a probability density through a sequence of *invertible*, *smooth* mappings to another one.

# The Basic Rule behind Normalizing Flows




$$
\begin{aligned}
\widehat{P}_{u_i}(\widehat{u}_i)\Delta\widehat{u}_i &= \widehat{P}_{s_i}(\widehat{s}_i)\Delta\widehat{s}_i \\
\widehat{P}_{u_i}(\widehat{u}_i) &= \frac{\Delta\widehat{s}_i}{\Delta\widehat{u}_i}\widehat{P}_{s_i}(\widehat{s}_i) \\
&= \frac{1}{|a|}\widehat{P}_{s_i}(\widehat{s}_i) \\
&= \frac{1}{\left|d\widehat{u}_i/d\widehat{s}_i\right|}\widehat{P}_{s_i}(\widehat{s}_i) \\
&= \left|\frac{d\widehat{s}_i}{d\widehat{u}_i}\right|\widehat{P}_{s_i}(\widehat{s}_i)
\end{aligned}
$$

Figure: Conservation of Probability Densities

# Finite Flows

Given:

- a random variable $\mathbf{z}_0$ with distribution $q_0$
- a chain of $K$ transformations $f_K$

The final random variable $\mathbf{z}_K$, and its density $q_K(\mathbf{z})$ is:

$$\mathbf{z}_K = f_K \circ \ldots \circ f_2 \circ f_1(\mathbf{z}_0) \tag{1}$$

$$\ln q_K(\mathbf{z}_K) = \ln q_0(\mathbf{z}_0) - \sum_{k=1}^{K} \ln \left| \det \frac{\partial f_k}{\partial \mathbf{z}_{k-1}} \right| \tag{2}$$

where (1) is called the *flow*, and (2) is a *normalizing flow*

# The Law of The Unconscious Statistician(LOTUS)

The previous normalizing flows has a nice analytic characteristic, is that expectations w.r.t. the transformed density $q_K$ can be calculated without knowing $q_K$:

$$\mathbb{E}_{q_K}[h(\mathbf{z})] = \mathbb{E}_{q_0}[h(f_K \circ ... \circ f_1(\mathbf{z}_0))] \tag{3}$$

It is unnecessary to calculate the log determinant of the Jacobian terms when $h(\mathbf{z})$ is not dependent on $q_K$.

# The Intuitive Effect of Invertible Flows

- Expansion: reducing the density in that region while increasing the density outside the region.
- Contraction: increasing the density in its interior while reducing the density outside.

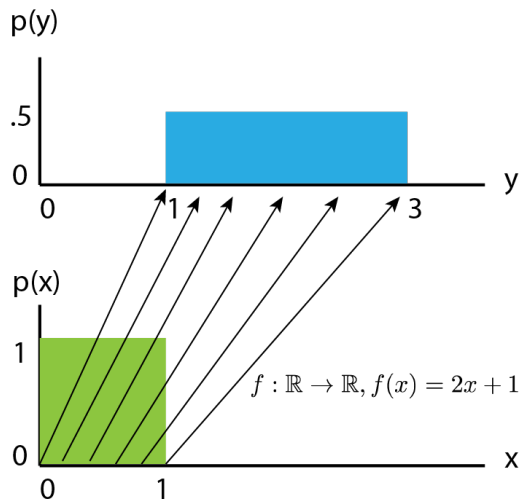



Figure: An example of expansion

# Approximating Posterior with Finite Flows

## Scheme:

Given:

- An simple factorized distributions $q_0$, e.g. independent Gaussian
- An appropriate choice of transformations $f_K$ with different lengths $K$

Obtain:

- $\rightsquigarrow$ increasingly complex and multi-model distributions

# Infinitesimal Flows

If the length of the normalizing flows tends to infinity, this kind of flows is no longer a sequence of transformations, but evolution of the initial density $q_0$ over time.

## Evolution Equation

$$\frac{\partial}{\partial t} q_t(\mathbf{z}) = \tau_t[q_t(\mathbf{z})]$$

where $\tau$ is the continuous-time dynamics. There are two typical

infinitesimal flows used:

- Langevin Flows based on Langevin dynamics, similar to Langevin Monte Carlo
- Hamiltonian Flows utilizing the mechanical energy balance in a closed system

# Invertible Linear-time Transformations

# Computational Cost of the Determinant

Using the following equation, an invertible parametric functions $q(\mathbf{z}')$ could be built:

$$q(\mathbf{z}') = q(\mathbf{z})|\det \frac{\partial f^{-1}{}'}{\partial \mathbf{z}}| = q(\mathbf{z})|\det \frac{\partial f}{\partial \mathbf{z}}|^{-1} \tag{4}$$

However, it's computational consumption:

- The Jacobian determinant $\rightsquigarrow \mathcal{O}(LD^3)$
- The gradients of the Jacobian determinant $\rightsquigarrow \mathcal{O}(LD^3)$

In the ideal case, our normalizing flows allows for low-cost computation of the determinant, or the Jacobian is not needed at all.

# Planar Flows

A family of transformations of the form:

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

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

where: $h(\cdot) = \tanh(\cdot)$ , $\psi(\cdot) = h'(\mathbf{w}^T \mathbf{z} + b)\mathbf{w}$ and all parameters $\lambda = \{\mathbf{w} \in \mathbb{R}^D, \mathbf{u} \in \mathbb{R}^D, b \in \mathbb{R}\}$

The resulting density $q_K(\mathbf{z})$ is implicitly given by:

$$\mathbf{z}_K = f_k \circ f_{K-1} \circ ... \circ f_1(\mathbf{z})$$

$$\ln q_K(\mathbf{z}_K) = \ln q_0(\mathbf{z}) - \sum_{k=1}^{K} \ln |1 + \mathbf{u}_k^T \psi_k(\mathbf{z}_{k-1})|$$

# Radial Flows

A family of transformations that modify an initial density $q_0$ around a reference point $\mathbf{z}_0$:

$$f(\mathbf{z}) = \mathbf{z} + \beta \cdot h(\alpha, r)(\mathbf{z} - \mathbf{z}_0)$$

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

where: $r = \|\mathbf{z} - \mathbf{z}_0\|$, $h(\alpha, r) = 1/(\alpha + r)$, $h'(\alpha, r) = -1/(\alpha + r)^2$ and All parameters: $\lambda = \{\mathbf{z}_0 \in \mathbb{R}^D, \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}\}$

The final density $q_K(\mathbf{z})$ is implicitly given by:

$$\mathbf{z}_K = f_k \circ f_{K-1} \circ ... \circ f_1(\mathbf{z})$$

$$\ln q_K(\mathbf{z}_K) = \ln q_0(\mathbf{z}) - \sum_{k=1}^{K} \ln |\det \frac{\partial f}{\partial \mathbf{z}_{k-1}}|$$

# Effects of Given Normalizing Flows

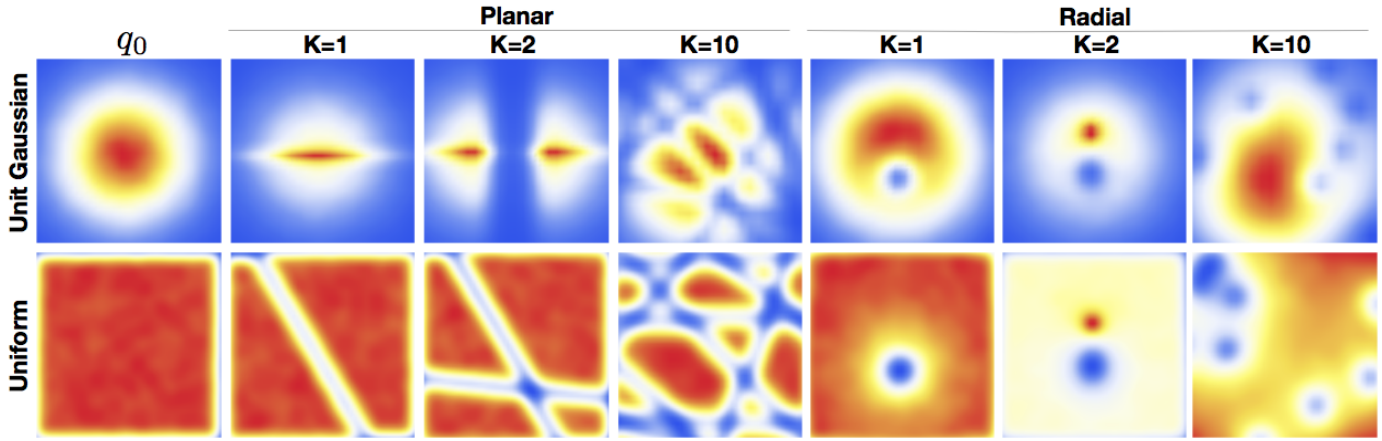



Figure: Effect of normalizing flow on two distributions

# Our Tasks

Given the formulas of the following four test energy functions with form $p(\mathbf{z}) \sim \exp(-U(\mathbf{z}))$:

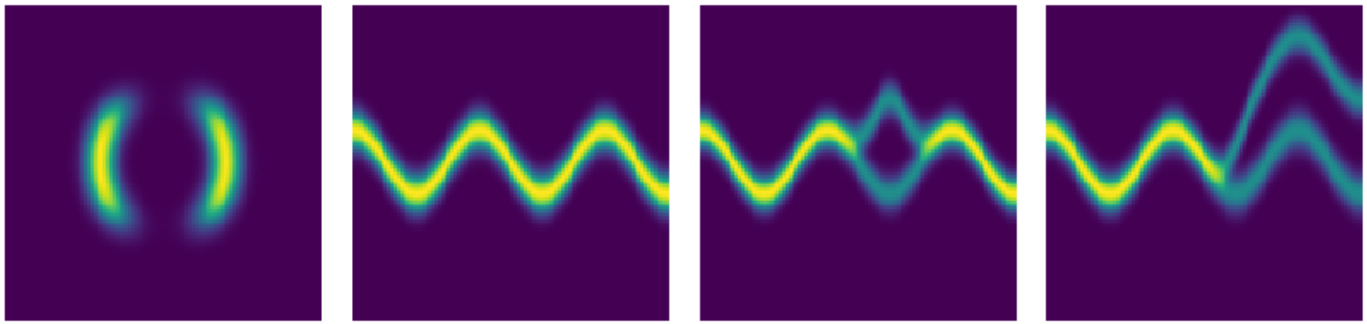


Figure: True posterior densities

Then, using deep neural networks $f(\cdot)$ to get $\mathbf{z}_0 \sim \mathcal{N}(0, \mathcal{I}) \overset{f_K \circ \ldots \circ f_1(\mathbf{z})}{\longrightarrow} p(\mathbf{z})$

# Framework of The Model

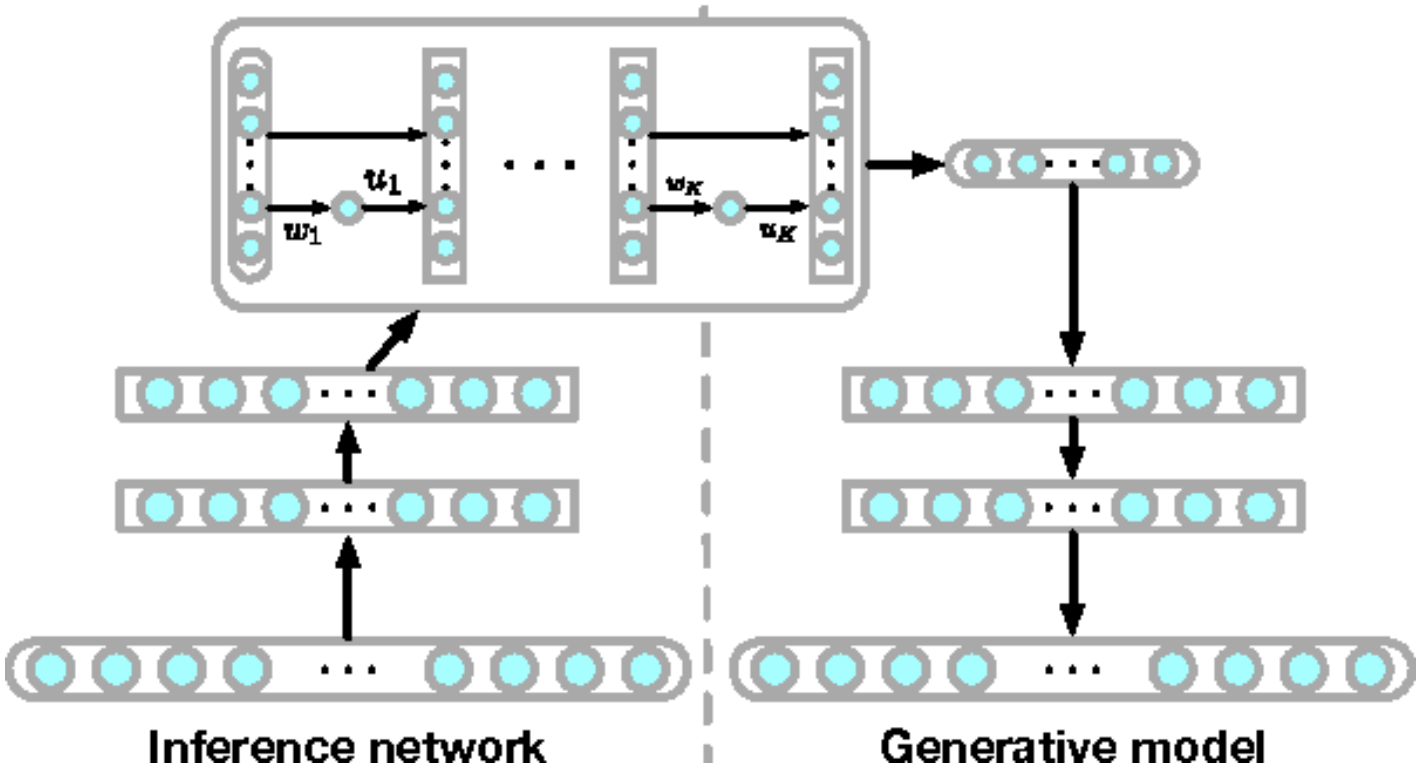



Figure: Inference and generative models

# Alternative Flow-based Posteriors

Now, we already have a method for linear-time computation of Jacobian. Could we find a method which the computation of Jacobian is not necessary?

The answer is: Yes

The Non-linear Independent Components Estimation(NICE):

$$|\det \frac{\partial f}{\partial \mathbf{z}}|^{-1} = 1$$

# Main Idea of NICE

## Factors in NICE

- $K$ neural networks $f(\cdot)$ with easy computed inverses $g(\cdot)$
- An arbitrary partitioning of $\mathbf{z} = (\mathbf{z}_A, \mathbf{z}_B)$
- Another neural networks $h_\lambda$ with parameters $\lambda$

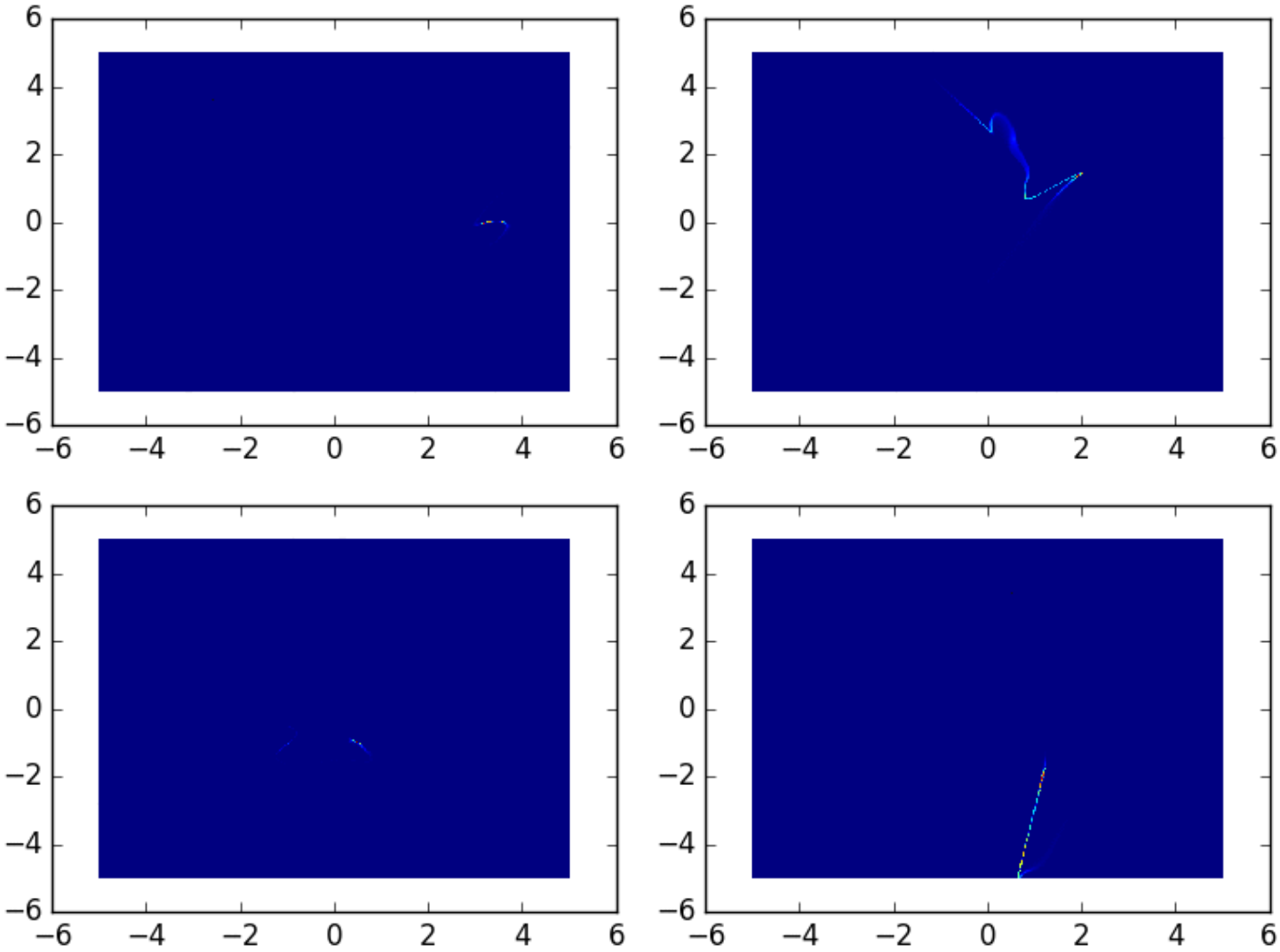
## Forward transformations

- $f(\mathbf{z}) = (\mathbf{z}_A, \mathbf{z}_B + h_\lambda(\mathbf{z}_A))$
- $\ln q_K(f_K \circ f_{K-1} \circ ... \circ f_1(\mathbf{z}_0)) = \ln q_0(\mathbf{z}_0)$

## Backward transformations

- $g(\mathbf{z}) = (\mathbf{z}_A, \mathbf{z}_B - h_\lambda(\mathbf{z}_A))$
- $\ln q_K(\mathbf{z}') = \ln q_0(g_1 \circ g_2 \circ ... \circ g_K(\mathbf{z})')$

# Power of Planar Flows



# Q&A

Thank you for your attention!