

# Deep Generative Models

## Lecture 2

Roman Isachenko



AI Masters

2024, Spring

## Recap of previous lecture

We are given i.i.d. samples  $\{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^m$  from unknown distribution  $\pi(\mathbf{x})$ .

### Goal

We would like to learn a distribution  $\pi(\mathbf{x})$  for

- ▶ evaluating  $\pi(\mathbf{x})$  for new samples (how likely to get object  $\mathbf{x}$ ?);
- ▶ sampling from  $\pi(\mathbf{x})$  (to get new objects  $\mathbf{x} \sim \pi(\mathbf{x})$ ).

Instead of searching true  $\pi(\mathbf{x})$  over all probability distributions, learn function approximation  $p(\mathbf{x}|\theta) \approx \pi(\mathbf{x})$ .

### Divergence

- ▶  $D(\pi||p) \geq 0$  for all  $\pi, p \in \mathcal{P}$ ;
- ▶  $D(\pi||p) = 0$  if and only if  $\pi \equiv p$ .

### Divergence minimization task

$$\min_{\theta} D(\pi||p).$$

# Recap of previous lecture

## Forward KL

$$KL(\pi||p) = \int \pi(\mathbf{x}) \log \frac{\pi(\mathbf{x})}{p(\mathbf{x}|\theta)} d\mathbf{x} \rightarrow \min_{\theta}$$

## Reverse KL

$$KL(p||\pi) = \int p(\mathbf{x}|\theta) \log \frac{p(\mathbf{x}|\theta)}{\pi(\mathbf{x})} d\mathbf{x} \rightarrow \min_{\theta}$$

## Maximum likelihood estimation (MLE)

$$\theta^* = \arg \max_{\theta} \prod_{i=1}^n p(\mathbf{x}_i|\theta) = \arg \max_{\theta} \sum_{i=1}^n \log p(\mathbf{x}_i|\theta).$$

Maximum likelihood estimation is equivalent to minimization of the Monte-Carlo estimate of forward KL.

# Recap of previous lecture

## Likelihood as product of conditionals

Let  $\mathbf{x} = (x_1, \dots, x_m)$ ,  $\mathbf{x}_{1:j} = (x_1, \dots, x_j)$ . Then

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^m p(x_j|\mathbf{x}_{1:j-1}, \boldsymbol{\theta}); \quad \log p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{j=1}^m \log p(x_j|\mathbf{x}_{1:j-1}, \boldsymbol{\theta}).$$

## MLE problem for autoregressive model

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^n \sum_{j=1}^m \log p(x_{ij}|\mathbf{x}_{i,1:j-1}\boldsymbol{\theta}).$$

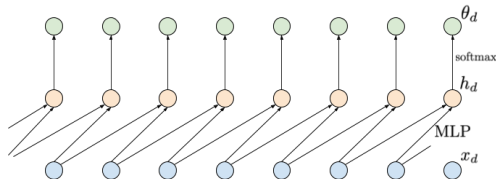
## Sampling

$$\hat{x}_1 \sim p(x_1|\boldsymbol{\theta}), \quad \hat{x}_2 \sim p(x_2|\hat{x}_1, \boldsymbol{\theta}), \quad \dots, \quad \hat{x}_m \sim p(x_m|\hat{\mathbf{x}}_{1:m-1}, \boldsymbol{\theta})$$

New generated object is  $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_m)$ .

# Recap of previous lecture

## Autoregressive MLP



## Autoregressive CNN

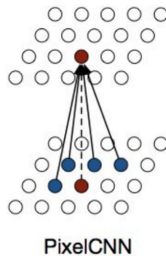
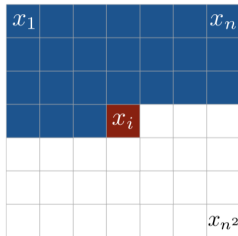


image credit: [https://jmtomczak.github.io/blog/2/2\\_ARM.html](https://jmtomczak.github.io/blog/2/2_ARM.html)

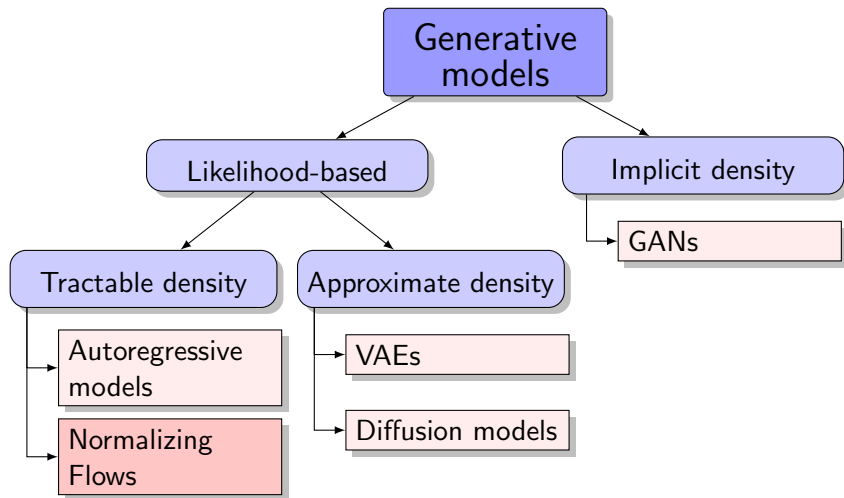
# Outline

1. Normalizing flows (NF)
2. Forward and Reverse KL for NF
3. NF examples
  - Linear normalizing flows
  - Gaussian autoregressive NF

# Outline

1. Normalizing flows (NF)
2. Forward and Reverse KL for NF
3. NF examples
  - Linear normalizing flows
  - Gaussian autoregressive NF

# Generative models zoo





# Normalizing flows prerequisites

## Jacobian matrix

Let  $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$  be a differentiable function.

$$\mathbf{z} = f(\mathbf{x}), \quad \mathbf{J} = \frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial z_1}{\partial x_1} & \cdots & \frac{\partial z_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_m}{\partial x_1} & \cdots & \frac{\partial z_m}{\partial x_m} \end{pmatrix} \in \mathbb{R}^{m \times m}$$

## Change of variable theorem (CoV)

Let  $\mathbf{x}$  be a random variable with density function  $p(\mathbf{x})$  and  $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$  is a differentiable, **invertible** function. If  $\mathbf{z} = f(\mathbf{x})$ ,  $\mathbf{x} = f^{-1}(\mathbf{z}) = g(\mathbf{z})$ , then

$$\begin{aligned} p(\mathbf{x}) &= p(\mathbf{z}) |\det(\mathbf{J}_f)| = p(\mathbf{z}) \left| \det \left( \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \right) \right| = p(f(\mathbf{x})) \left| \det \left( \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right) \right| \\ p(\mathbf{z}) &= p(\mathbf{x}) |\det(\mathbf{J}_g)| = p(\mathbf{x}) \left| \det \left( \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \right) \right| = p(g(\mathbf{z})) \left| \det \left( \frac{\partial g(\mathbf{z})}{\partial \mathbf{z}} \right) \right|. \end{aligned}$$

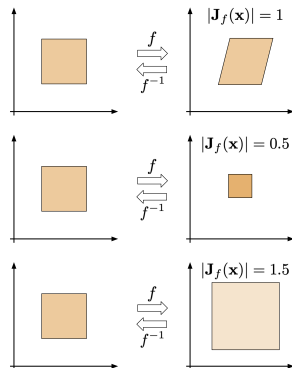
# Jacobian determinant

## Inverse function theorem

If function  $f$  is invertible and Jacobian matrix is continuous and non-singular, then

$$\mathbf{J}_{f^{-1}} = \mathbf{J}_g = \mathbf{J}_f^{-1}; \quad |\det(\mathbf{J}_{f^{-1}})| = |\det(\mathbf{J}_g)| = \frac{1}{|\det(\mathbf{J}_f)|}.$$

- ▶  $\mathbf{x}$  and  $\mathbf{z}$  have the same dimensionality ( $\mathbb{R}^m$ ).
- ▶  $f_\theta(\mathbf{x})$  could be parametric function.
- ▶ Determinant of Jacobian matrix  $\mathbf{J} = \frac{\partial f_\theta(\mathbf{x})}{\partial \mathbf{x}}$  shows how the volume changes under the transformation.



# Fitting normalizing flows

## MLE problem

$$p(\mathbf{x}|\theta) = p(\mathbf{z}) \left| \det \left( \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \right) \right| = p(f_{\theta}(\mathbf{x})) \left| \det \left( \frac{\partial f_{\theta}(\mathbf{x})}{\partial \mathbf{x}} \right) \right|$$

$$\log p(\mathbf{x}|\theta) = \log p(f_{\theta}(\mathbf{x})) + \log |\det(\mathbf{J}_f)| \rightarrow \max_{\theta}$$

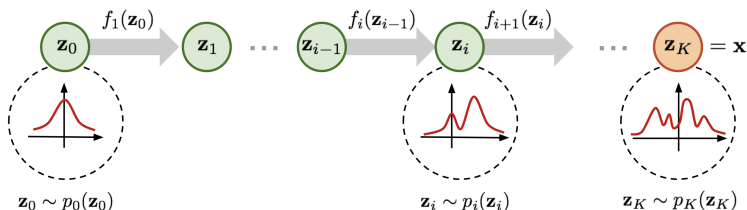


# Composition of normalizing flows

## Theorem

If  $\{f_k\}_{k=1}^K$  satisfy conditions of the change of variable theorem, then  $\mathbf{z} = f(\mathbf{x}) = f_K \circ \dots \circ f_1(\mathbf{x})$  also satisfies it.

$$\begin{aligned} p(\mathbf{x}) &= p(f(\mathbf{x})) \left| \det \left( \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right) \right| = p(f(\mathbf{x})) \left| \det \left( \frac{\partial \mathbf{f}_K}{\partial \mathbf{f}_{K-1}} \dots \frac{\partial \mathbf{f}_1}{\partial \mathbf{x}} \right) \right| = \\ &= p(f(\mathbf{x})) \prod_{k=1}^K \left| \det \left( \frac{\partial \mathbf{f}_k}{\partial \mathbf{f}_{k-1}} \right) \right| = p(f(\mathbf{x})) \prod_{k=1}^K |\det(\mathbf{J}_{f_k})| \end{aligned}$$



# Normalizing flows (NF)

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log p(f_{\boldsymbol{\theta}}(\mathbf{x})) + \log |\det(\mathbf{J}_f)|$$

## Definition

Normalizing flow is a *differentiable, invertible* mapping from data  $\mathbf{x}$  to the noise  $\mathbf{z}$ .

- ▶ **Normalizing** means that NF takes samples from  $\pi(\mathbf{x})$  and normalizes them into samples from the density  $p(\mathbf{z})$ .
- ▶ **Flow** refers to the trajectory followed by samples from  $p(\mathbf{z})$  as they are transformed by the sequence of transformations

$$\mathbf{z} = f_K \circ \dots \circ f_1(\mathbf{x}); \quad \mathbf{x} = f_1^{-1} \circ \dots \circ f_K^{-1}(\mathbf{z}) = g_1 \circ \dots \circ g_K(\mathbf{z})$$

## Log likelihood

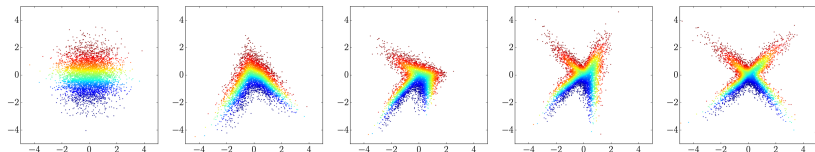
$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log p(f_K \circ \dots \circ f_1(\mathbf{x})) + \sum_{k=1}^K \log |\det(\mathbf{J}_{f_k})|,$$

where  $\mathbf{J}_{f_k} = \frac{\partial \mathbf{f}_k}{\partial \mathbf{f}_{k-1}}$ .

**Note:** Here we consider only **continuous** random variables.

# Normalizing flows

## Example of a 4-step NF



## NF log likelihood

$$\log p(\mathbf{x}|\theta) = \log p(f_{\theta}(\mathbf{x})) + \log |\det(\mathbf{J}_f)|$$

What is the complexity of the determinant computation?

What do we need?

- ▶ efficient computation of the Jacobian matrix  $\mathbf{J}_f = \frac{\partial f_{\theta}(\mathbf{x})}{\partial \mathbf{x}}$ ;
- ▶ efficient inversion of  $f_{\theta}(\mathbf{x})$ .

# Outline

1. Normalizing flows (NF)
2. Forward and Reverse KL for NF
3. NF examples
  - Linear normalizing flows
  - Gaussian autoregressive NF

# Forward KL vs Reverse KL

Forward KL  $\equiv$  MLE

$$\begin{aligned} KL(\pi||p) &= \int \pi(\mathbf{x}) \log \frac{\pi(\mathbf{x})}{p(\mathbf{x}|\boldsymbol{\theta})} d\mathbf{x} \\ &= -\mathbb{E}_{\pi(\mathbf{x})} \log p(\mathbf{x}|\boldsymbol{\theta}) + \text{const} \rightarrow \min_{\boldsymbol{\theta}} \end{aligned}$$

Forward KL for NF model

$$\begin{aligned} \log p(\mathbf{x}|\boldsymbol{\theta}) &= \log p(f_{\boldsymbol{\theta}}(\mathbf{x})) + \log |\det(\mathbf{J}_f)| \\ KL(\pi||p) &= -\mathbb{E}_{\pi(\mathbf{x})} [\log p(f_{\boldsymbol{\theta}}(\mathbf{x})) + \log |\det(\mathbf{J}_f)|] + \text{const} \end{aligned}$$

- ▶ We need to be able to compute  $f_{\boldsymbol{\theta}}(\mathbf{x})$  and its Jacobian.
- ▶ We need to be able to compute the density  $p(\mathbf{z})$ .
- ▶ We don't need to think about computing the function  $g_{\boldsymbol{\theta}}(\mathbf{z}) = f^{-1}(\mathbf{z}, \boldsymbol{\theta})$  until we want to sample from the NF.



# Forward KL vs Reverse KL

## Reverse KL

$$\begin{aligned} KL(p||\pi) &= \int p(\mathbf{x}|\boldsymbol{\theta}) \log \frac{p(\mathbf{x}|\boldsymbol{\theta})}{\pi(\mathbf{x})} d\mathbf{x} \\ &= \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})} [\log p(\mathbf{x}|\boldsymbol{\theta}) - \log \pi(\mathbf{x})] \rightarrow \min_{\boldsymbol{\theta}} \end{aligned}$$

## Reverse KL for NF model (LOTUS trick)

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log p(\mathbf{z}) + \log |\det(\mathbf{J}_f)| = \log p(\mathbf{z}) - \log |\det(\mathbf{J}_g)|$$

$$KL(p||\pi) = \mathbb{E}_{p(\mathbf{z})} [\log p(\mathbf{z}) - \log |\det(\mathbf{J}_g)| - \log \pi(g_{\boldsymbol{\theta}}(\mathbf{z}))]$$

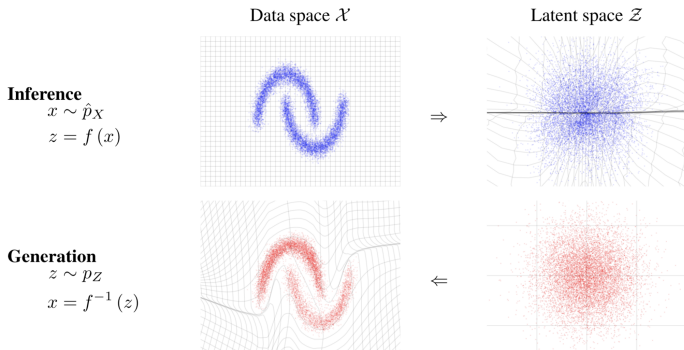
- ▶ We need to be able to compute  $g_{\boldsymbol{\theta}}(\mathbf{z})$  and its Jacobian.
- ▶ We need to be able to sample from the density  $p(\mathbf{z})$  (do not need to evaluate it) and to evaluate(!)  $\pi(\mathbf{x})$ .
- ▶ We don't need to think about computing the function  $f_{\boldsymbol{\theta}}(\mathbf{x})$ .

# Normalizing flows KL duality

## Theorem

Fitting NF model  $p(\mathbf{x}|\boldsymbol{\theta})$  to the target distribution  $\pi(\mathbf{x})$  using forward KL (MLE) is equivalent to fitting the induced distribution  $p(\mathbf{z}|\boldsymbol{\theta})$  to the base  $p(\mathbf{z})$  using reverse KL:

$$\arg \min_{\boldsymbol{\theta}} KL(\pi(\mathbf{x})||p(\mathbf{x}|\boldsymbol{\theta})) = \arg \min_{\boldsymbol{\theta}} KL(p(\mathbf{z}|\boldsymbol{\theta})||p(\mathbf{z})).$$



# Normalizing flows KL duality

## Theorem

$$\arg \min_{\theta} KL(\pi(\mathbf{x})||p(\mathbf{x}|\theta)) = \arg \min_{\theta} KL(p(\mathbf{z}|\theta)||p(\mathbf{z})).$$

## Proof

- ▶  $\mathbf{z} \sim p(\mathbf{z}), \mathbf{x} = g_{\theta}(\mathbf{z}), \mathbf{x} \sim p(\mathbf{x}|\theta);$
- ▶  $\mathbf{x} \sim \pi(\mathbf{x}), \mathbf{z} = f_{\theta}(\mathbf{x}), \mathbf{z} \sim p(\mathbf{z}|\theta);$

$$\log p(\mathbf{z}|\theta) = \log \pi(g_{\theta}(\mathbf{z})) + \log |\det(\mathbf{J}_g)|;$$

$$\log p(\mathbf{x}|\theta) = \log p(f_{\theta}(\mathbf{x})) + \log |\det(\mathbf{J}_f)|.$$

$$\begin{aligned} KL(p(\mathbf{z}|\theta)||p(\mathbf{z})) &= \mathbb{E}_{p(\mathbf{z}|\theta)} [\log p(\mathbf{z}|\theta) - \log p(\mathbf{z})] = \\ &= \mathbb{E}_{p(\mathbf{z}|\theta)} [\log \pi(g_{\theta}(\mathbf{z})) + \log |\det(\mathbf{J}_g)| - \log p(\mathbf{z})] = \\ &= \mathbb{E}_{\pi(\mathbf{x})} [\log \pi(\mathbf{x}) - \log |\det(\mathbf{J}_f)| - \log p(f_{\theta}(\mathbf{x}))] = \\ &= \mathbb{E}_{\pi(\mathbf{x})} [\log \pi(\mathbf{x}) - \log p(\mathbf{x}|\theta)] = KL(\pi(\mathbf{x})||p(\mathbf{x}|\theta)). \end{aligned}$$

# Outline

1. Normalizing flows (NF)
2. Forward and Reverse KL for NF
3. NF examples
  - Linear normalizing flows
  - Gaussian autoregressive NF

# Outline

1. Normalizing flows (NF)
2. Forward and Reverse KL for NF
3. NF examples
  - Linear normalizing flows
  - Gaussian autoregressive NF

# Jacobian structure

## Normalizing flows log-likelihood

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log p(f_{\boldsymbol{\theta}}(\mathbf{x})) + \log \left| \det \left( \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \mathbf{x}} \right) \right|$$

The main challenge is a determinant of the Jacobian matrix.

What is the  $\det(\mathbf{J})$  in the following cases?

Consider a linear layer  $\mathbf{z} = \mathbf{W}\mathbf{x}$ ,  $\mathbf{W} \in \mathbb{R}^{m \times m}$ .

1. Let  $\mathbf{z}$  be a permutation of  $\mathbf{x}$ .
2. Let  $z_j$  depend only on  $x_j$ .

$$\log \left| \det \left( \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \mathbf{x}} \right) \right| = \log \left| \prod_{j=1}^m \frac{\partial f_{j,\boldsymbol{\theta}}(x_j)}{\partial x_j} \right| = \sum_{j=1}^m \log \left| \frac{\partial f_{j,\boldsymbol{\theta}}(x_j)}{\partial x_j} \right|.$$

3. Let  $z_j$  depend only on  $\mathbf{x}_{1:j}$  (autoregressive dependency).

# Linear normalizing flows

$$\mathbf{z} = f_{\theta}(\mathbf{x}) = \mathbf{W}\mathbf{x}, \quad \mathbf{W} \in \mathbb{R}^{m \times m}, \quad \theta = \mathbf{W}, \quad \mathbf{J}_f = \mathbf{W}^T$$

In general, we need  $O(m^3)$  to invert matrix.

## Invertibility

- ▶ Diagonal matrix  $O(m)$ .
- ▶ Triangular matrix  $O(m^2)$ .
- ▶ It is impossible to parametrize all invertible matrices.

## Invertible 1x1 conv

$\mathbf{W} \in \mathbb{R}^{c \times c}$  – kernel of 1x1 convolution with  $c$  input and  $c$  output channels. The computational complexity of computing or differentiating  $\det(\mathbf{W})$  is  $O(c^3)$ . Cost to compute  $\det(\mathbf{W})$  is  $O(c^3)$ . It should be invertible.

# Linear normalizing flows

$$\mathbf{z} = f_{\theta}(\mathbf{x}) = \mathbf{W}\mathbf{x}, \quad \mathbf{W} \in \mathbb{R}^{m \times m}, \quad \theta = \mathbf{W}, \quad \mathbf{J}_f = \mathbf{W}^T$$

## Matrix decompositions

### ► LU-decomposition

$$\mathbf{W} = \mathbf{P}\mathbf{L}\mathbf{U},$$

where  $\mathbf{P}$  is a permutation matrix,  $\mathbf{L}$  is lower triangular with positive diagonal,  $\mathbf{U}$  is upper triangular with positive diagonal.

### ► QR-decomposition

$$\mathbf{W} = \mathbf{Q}\mathbf{R},$$

where  $\mathbf{Q}$  is an orthogonal matrix,  $\mathbf{R}$  is an upper triangular matrix with positive diagonal.

Decomposition should be done only once in the beginning. Next, we fit decomposed matrices ( $\mathbf{P}/\mathbf{L}/\mathbf{U}$  or  $\mathbf{Q}/\mathbf{R}$ ).

---

*Kingma D. P., Dhariwal P. Glow: Generative Flow with Invertible 1x1 Convolutions, 2018*

*Hoogeboom E., et al. Emerging convolutions for generative normalizing flows, 2019*



# Outline

1. Normalizing flows (NF)
2. Forward and Reverse KL for NF
3. NF examples
  - Linear normalizing flows
  - Gaussian autoregressive NF

# Gaussian autoregressive model

Consider an autoregressive model

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^m p(x_j|\mathbf{x}_{1:j-1}, \boldsymbol{\theta}), \quad p(x_j|\mathbf{x}_{1:j-1}, \boldsymbol{\theta}) = \mathcal{N}(\mu_j(\mathbf{x}_{1:j-1}), \sigma_j^2(\mathbf{x}_{1:j-1})).$$

## Sampling

$$x_j = \sigma_j(\mathbf{x}_{1:j-1}) \cdot z_j + \mu_j(\mathbf{x}_{1:j-1}), \quad z_j \sim \mathcal{N}(0, 1).$$

## Inverse transform

$$z_j = (x_j - \mu_j(\mathbf{x}_{1:j-1})) \cdot \frac{1}{\sigma_j(\mathbf{x}_{1:j-1})}.$$

- ▶ We have an **invertible** and **differentiable** transformation from  $p(\mathbf{z})$  to  $p(\mathbf{x}|\boldsymbol{\theta})$ .
- ▶ It is an autoregressive (AR) NF with the base distribution  $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$ !
- ▶ Jacobian of such transformation is triangular!

# Gaussian autoregressive NF

$$\mathbf{x} = g_{\theta}(\mathbf{z}) \quad \Rightarrow \quad x_j = \sigma_j(\mathbf{x}_{1:j-1}) \cdot z_j + \mu_j(\mathbf{x}_{1:j-1}).$$

$$\mathbf{z} = f_{\theta}(\mathbf{x}) \quad \Rightarrow \quad z_j = (x_j - \mu_j(\mathbf{x}_{1:j-1})) \cdot \frac{1}{\sigma_j(\mathbf{x}_{1:j-1})}.$$

Generation function  $g_{\theta}(\mathbf{z})$  is **sequential**.

Inference function  $f_{\theta}(\mathbf{x})$  is **not sequential**.

## Forward KL for NF

$$KL(\pi||p) = -\mathbb{E}_{\pi(\mathbf{x})} [\log p(f_{\theta}(\mathbf{x})) + \log |\det(\mathbf{J}_f)|] + \text{const}$$

- ▶ We need to be able to compute  $f_{\theta}(\mathbf{x})$  and its Jacobian.
- ▶ We need to be able to compute the density  $p(\mathbf{z})$ .
- ▶ We don't need to think about computing the function  $g_{\theta}(\mathbf{z}) = f_{\theta}^{-1}(\mathbf{z})$  until we want to sample from the model.

# Gaussian autoregressive NF

$$\mathbf{x} = g_{\theta}(\mathbf{z}) \Rightarrow x_j = \sigma_j(\mathbf{x}_{1:j-1}) \cdot z_j + \mu_j(\mathbf{x}_{1:j-1}).$$

$$\mathbf{z} = f_{\theta}(\mathbf{x}) \Rightarrow z_j = (x_j - \mu_j(\mathbf{x}_{1:j-1})) \cdot \frac{1}{\sigma_j(\mathbf{x}_{1:j-1})}.$$

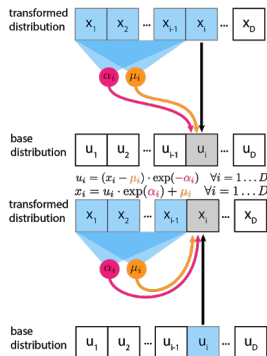
- ▶ Sampling is sequential, density estimation is parallel.
- ▶ Forward KL is a natural loss.

Forward transform:  $f_{\theta}(\mathbf{x})$

$$z_j = (x_j - \mu_j(\mathbf{x}_{1:j-1})) \cdot \frac{1}{\sigma_j(\mathbf{x}_{1:j-1})}$$

Inverse transform:  $g_{\theta}(\mathbf{z})$

$$x_j = \sigma_j(\mathbf{x}_{1:j-1}) \cdot z_j + \mu_j(\mathbf{x}_{1:j-1})$$



# Summary

- ▶ Change of variable theorem allows to get the density function of the random variable under the invertible transformation.
- ▶ Normalizing flows transform a simple base distribution to a complex one via a sequence of invertible transformations with tractable Jacobian.
- ▶ Normalizing flows have a tractable likelihood that is given by the change of variable theorem.
- ▶ We fit normalizing flows using forward or reverse KL minimization.
- ▶ Linear NF try to parametrize set of invertible matrices via matrix decompositions.
- ▶ Gaussian autoregressive NF is an autoregressive model with triangular Jacobian. It has fast inference function and slow generation function. Forward KL is a natural loss function.