Advanced Machine Learning:
Deep Generative Models

Normalizing Flows

lecturer: Prof. Dr. Stephan Günnemann
www.daml.in.tum.de

Summer Term 23

Data Analytics and Machine Learning

Roadmap

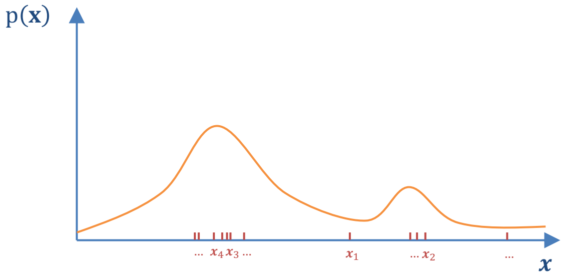
- Deep Generative Models
  1. Introduction
  2. **Normalizing Flows**
     - **Change of Variables Formula**
     - Forward and Reverse Parametrization
     - Jacobian Determinant Computation
  3. Variational Inference
  4. Variational Autoencoder
  5. Generative Adversarial Networks
  6. Denoising Diffusion

Motivation

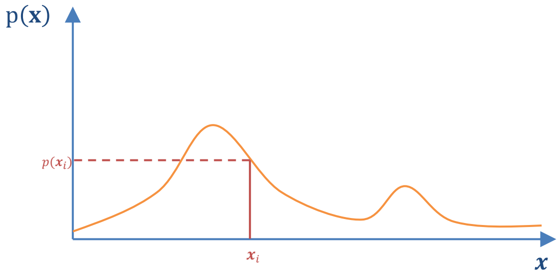
- We assume that the data $\boldsymbol{x}$ follows a probability distribution $\mathrm{p}(\boldsymbol{x})$ i.e.

$$\mathrm{p}(\boldsymbol{x}) \text{ where } \boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix}$$

- We can do two interesting things with a distribution:



Data sampling: generate data sample $\boldsymbol{x}_i$ following the distribution $\mathrm{p}(\boldsymbol{x})$



Density evaluation: given any $\boldsymbol{x}_i$, compute the probability density at this point $\mathrm{p}(\boldsymbol{x}_i)$

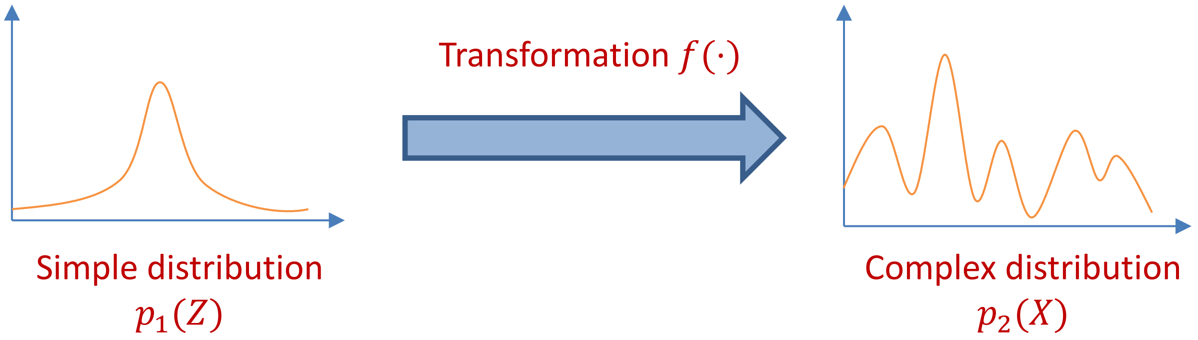
Motivation

- Normalizing Flows (NF) can model flexible distributions for *data sampling* and *density evaluation*.

- Normalizing Flows intuition:

Model a complex distribution

by applying a transformation on a simple distribution

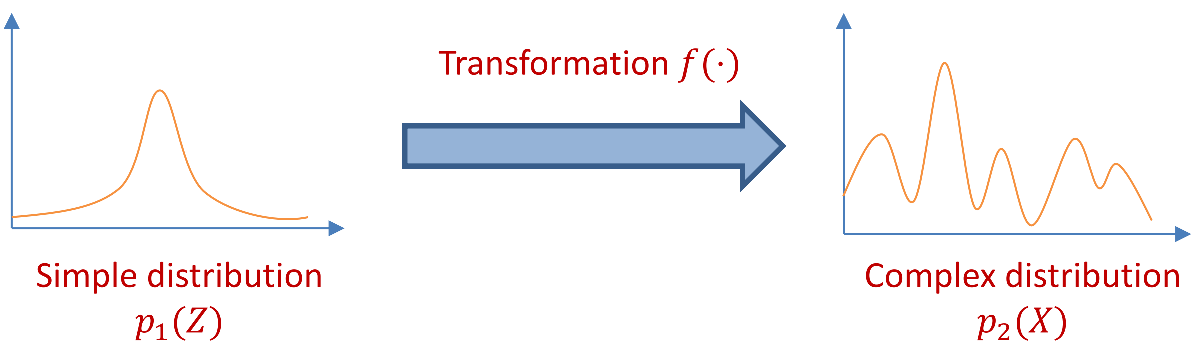


Idea

- Normalizing Flows are based on the change of variables formula

- It transforms a variable $\boldsymbol{z}$ into another variable $\boldsymbol{x}$ by via a transformation f i.e. $f(\boldsymbol{z}) = \boldsymbol{x}$

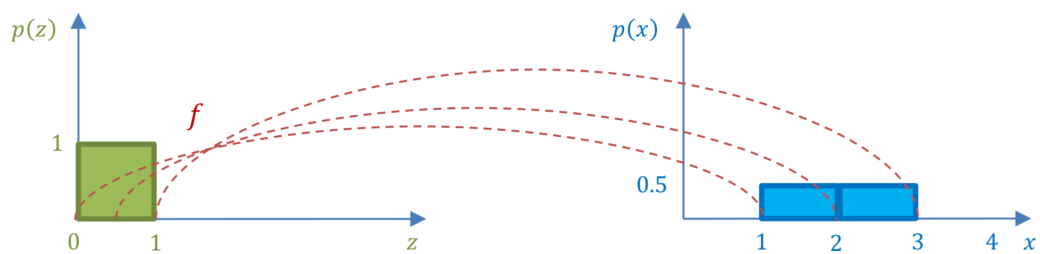
- It is particularly useful to simplify computations when working with distributions (or integrals)



Change of Variables: Example

Introductory example: $D = 1, p_1(z) = Unif([0,1]), f(z) = 2z + 1 = x$

- The probability in the input space should be the same as in the output space i.e. $p_1(z)\partial z = p_2(x)\partial x$

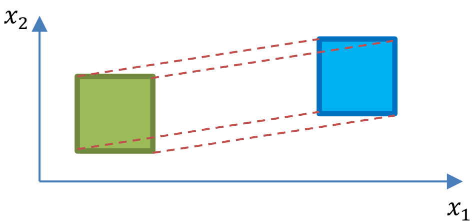


- Abusing the notations leads to $p_2(x) = p_1(z)\left|\frac{\partial z}{\partial x}\right|$
  - The term $\left|\frac{\partial z}{\partial x}\right|$ renormalize the probability distribution in the output space
  - The normalization term is the same if the transformation is increasing or decreasing

Change of Variables: Example

Introductory example: $D = 2, p_1(\boldsymbol{z}) = Unif([0,1]^2), f(\boldsymbol{z}) = \boldsymbol{z} + shift = \boldsymbol{x}$

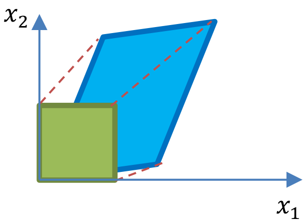
- Applying a constant shift does not change the area after the transformation i.e. $p_2\,(\boldsymbol{x}) = p_1(\boldsymbol{z})$



Introductory example: $D = 2, p_1(\boldsymbol{z}) = Unif([0,1]^2), f(\boldsymbol{z}) = M\boldsymbol{z} = \boldsymbol{x}, M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

- The linear transformation M changes the area from 1 to $ad - bc = \det(M)$.
- The probability distribution $p_2(\boldsymbol{x})$ has to be normalized

  i.e. $p_2\,(\boldsymbol{x}) = p_1(\boldsymbol{z})\ \frac{1}{det(M)}$



Change of Variables Formula

Change of variables formula (General case): if $D \in \mathbb{N}$, $\mathrm{p}_1(\boldsymbol{z})$ a D -dimensional distribution, $f(\boldsymbol{z}) = \boldsymbol{x}$ an *invertible* and *differentiable* transformation, then distribution $p_2(\boldsymbol{x})$ is

$$p_2(\boldsymbol{x}) = p_1(f^{-1}(\boldsymbol{x})) \cdot \left|\det\left(\frac{\partial f^{-1}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\right|$$

- The determinant term accounts for the **distortion rate** of the transformation (see introductory examples). If it is equal to 1, $p_2(\boldsymbol{x})$ and $p_1(\boldsymbol{z})$ have the same « volume » at this point i.e. $p_2(\boldsymbol{x}) = p_1(f^{-1}(\boldsymbol{x}))$.
  - It considers that the transformation is locally linear (see last example)

- The term $\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}}$ is called **Jacobian** of g; here: a $D \times D$ matrix
  - We have $\frac{\partial f^{-1}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left(\frac{\partial f(\boldsymbol{z})}{\partial \boldsymbol{z}}\right)^{-1}$.
- The transformation f should be ***valid*** *(invertible* and *differentiable).*

Conditions

(Sufficient) Conditions for a valid transformation f:

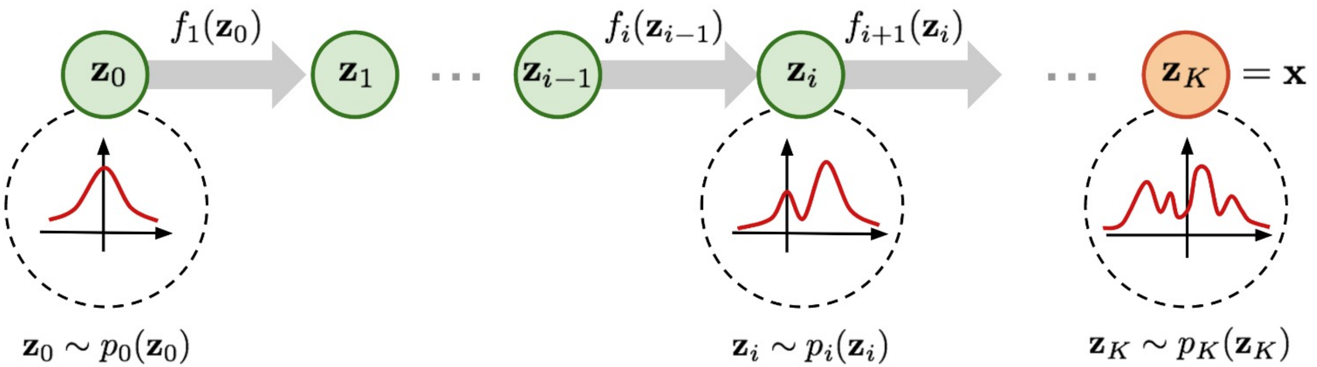
- *Invertibility:*
  - The input and output space of the mapping should have the **same dimension** $\boldsymbol{D}$.
  - If $\boldsymbol{D} = \mathbf{1}$, it is sufficient that f is **strictly monotonic** (increasing or decreasing).
  - If the transformation f is **linear**, its determinant should be nonzero i.e. $\mathbf{det}(\boldsymbol{f}) \neq \mathbf{0}$

- *Differentiability:*
  - Both the transformation f and its inverse f^{-1} are continuously differentiable, i.e. the Jacobian $\frac{\partial f^{-1}(\boldsymbol{x})}{\partial \boldsymbol{x}}$ exists at any point $\boldsymbol{x}$.
  - Note: Differentiability is a sufficient condition; in theory, the mapping f does not have to be differentiable everywhere, we can even have discontinuous; in practice we usually use only differentiable transformation

Stacking

Stacking transformations f_i:

- We can apply the change of variables formula multiple times:
  - The first variable $\boldsymbol{z_0}$ is transformed by $\boldsymbol{z_1} = f_1(\boldsymbol{z_0})$
  - The variable $\boldsymbol{z_{i-1}}$ is transformed by $\boldsymbol{z_i} = f_i(\boldsymbol{z_{i-1}})$
  - The last variable $\boldsymbol{z_{K-1}}$ is transformed by $\mathbf{x} = \boldsymbol{z_K} = f_K(\boldsymbol{z_{K-1}})$



- The change of variables formula becomes:

[Lilian Weng blog]

$$p_K(\boldsymbol{x}) = p_0(\boldsymbol{z_0}) \prod_{i=1}^{K} \left| \det\left(\frac{\partial f_i^{-1}(\boldsymbol{z_i})}{\partial \boldsymbol{z_i}} \right) \right|$$

Change of Variables Formula: Log Version

Change of variables formula (log version): if $D \in \mathbb{N}$, $p_1(\boldsymbol{z})$ a D-dim. distribution, $f(\boldsymbol{z}) = \boldsymbol{x}$ an *invertible* and *differentiable* transformation, then $p_2(\boldsymbol{x})$ is

$$\log(p_2(\boldsymbol{x})) = \log(p_1(f^{-1}(\boldsymbol{x}))) + \log\left|\det\left(\frac{\partial f^{-1}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\right|$$

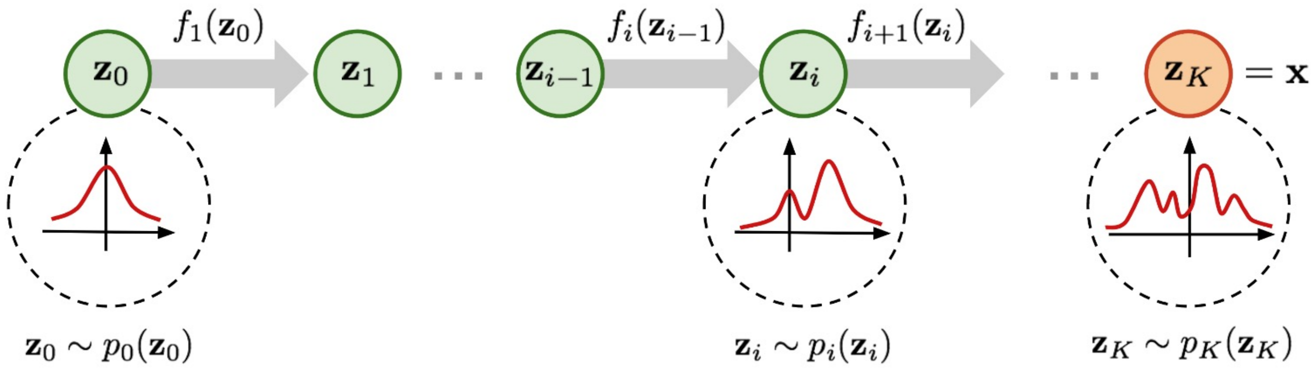
- Optimization is easier when working with sums. Consequently, we generally consider log probabilities (e.g. maximize log likelihood)

- The log version when stacking transformations is:

$$\log(p_K(\boldsymbol{x})) = \log(p_0(\boldsymbol{z_0})) + \sum_{i=1}^{K} \log\left|\det\left(\frac{\partial f_i^{-1}(\boldsymbol{z}_i)}{\partial \boldsymbol{z}_i}\right)\right|$$

The name: "Normalizing flow"

- A NF transforms a simple distribution (e.g. uniform, Gaussian) in a complex distribution. For some Normalizing Flows the universality theorem has been proven.
- NFs stack valid transformations to model a complex mapping between the input and output space. The input variable **flows** through the transformations.
- The change of variable formula allows to compute the distribution of the output space based on the distribution in the input space. The determinant terms **normalize** the distribution in the output space.



Questions – NF1

1. Is $\mathrm{f}(z) = 1 - \mathrm{z}$ a valid normalizing flow transformation?

2. Is $\mathrm{f}(z) = 2 - 3\,z$ a valid NF transformation?

3. Is $\mathrm{f}(z) = \begin{cases} -z, z \in [0,1] \\ 1 - z, z \in [1,2] \end{cases}$ a valid NF transformation?

Roadmap

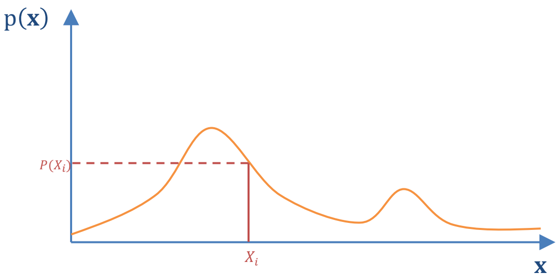
- Deep Generative Models
  1. Introduction
  2. **Normalizing Flows**
     – Change of Variables Formula
     – **Forward and Reverse Parametrization**
     – Jacobian Determinant Computation
  3. Variational Inference
  4. Variational Autoencoder
  5. Generative Adversarial Networks
  6. Denoising Diffusion

Forward and Reverse Parametrization

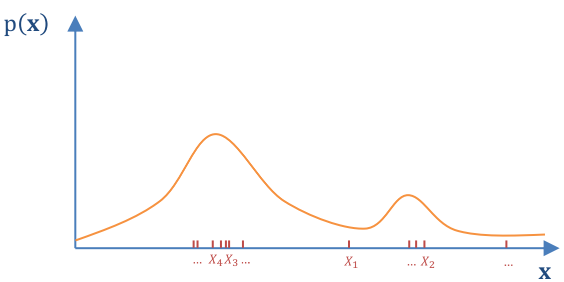
- The change of variables formula does a mapping between two distributions

$$p_2(\boldsymbol{x}) = p_1(f^{-1}(\boldsymbol{x})) \cdot \left|\det\left(\frac{\partial f^{-1}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\right|$$

- How to use the change of variables formula to evaluate $p_2(\boldsymbol{x})$ at any point $\boldsymbol{x}$?
  - ➢ **Reverse parametrization**



- How to use the change of variables formula to sample points from $p_2(\boldsymbol{x})$?
  - ➢ **Forward parametrization**



Forward and Reverse Parametrization

- There exist many different flows (see [3] for further details):
  - Planar/Radial flows
  - RealNVP
  - IAF
  - MAF
  - Spline
  - …

  Autoregressive Flows

- These flows generally differ in the parametrization of the transformation f.
  - Some parametrizations are efficient for sampling.
  - Some parametrizations are efficient for density evaluation.

- Efficient sampling and density evaluation (at the same time) might not be required in all applications

Reverse Parametrization

- We parametrize the inverse transformation $g = f^{-1}$ that we know analytically.
  - $g_\varphi(\boldsymbol{x}) = \boldsymbol{z}$ is computable and parameter φ can be learned

- We know that the inverse $f = g^{-1}$ exists, but we might not know it analytically.
  - $g_\varphi^{-1}(\boldsymbol{z}) = \boldsymbol{x}$ might not be (easily) computable

- The change of variable formula with reverse parametrization is

$$p_2(\boldsymbol{x}) = p_1(g_\varphi(\boldsymbol{x})) \cdot \left|\det\left(\frac{\partial g_\varphi(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\right|$$

  - The formula only uses the known parametrized function g_φ.
  - Given any point $\boldsymbol{x}^{(j)}$, we can compute $p_2(\boldsymbol{x}^{(j)})$.

Reverse Parametrization

Stacking:

- We can also stack transformations with reverse parametrization i.e.

$$g_{\varphi} = g_{\varphi_1} \circ \cdots \circ g_{\varphi_K}$$

- To **compute the density**
  1. For any $\boldsymbol{x}^{(j)}$, we can set $\boldsymbol{x}^{(j)} = \boldsymbol{z_K}$
  2. Compute the transformations $\boldsymbol{z}_{i-1}^{(j)} = g_{\varphi_i}(\boldsymbol{z}_i^{(j)})$ and $\left|\det\left(\frac{\partial g_{\varphi_i}\left(\boldsymbol{z_i}^{(j)}\right)}{\partial \boldsymbol{z_i}^{(j)}}\right)\right|$
  3. Given $\boldsymbol{z}_0^{(j)}$, we can compute $p_0(\boldsymbol{z}_0^{(j)})$ (e.g. Gaussian or Uniform) and thus $p_K(\boldsymbol{x}^{(j)})$

- Important reminder: While $p_0(\boldsymbol{z})$ has a simple shape, $p_K(\boldsymbol{x})$ can capture very complex structure
  - This is all realized by g_{φ}

Normalizing Flows in Machine Learning

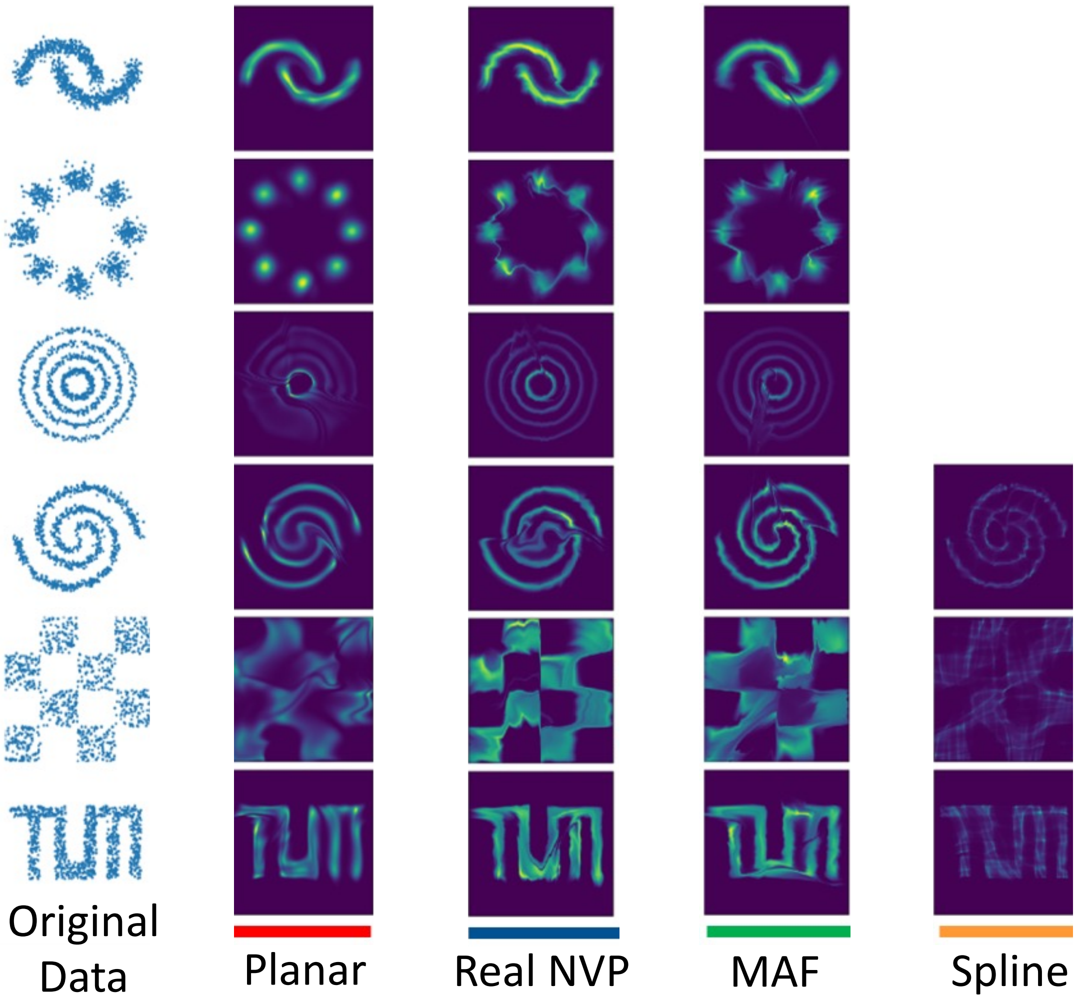
- Why do we care about computing the density $p_K(x^{(j)})$?
- We can use it for learning!
  - We aim to learn g_φ , i.e. the transformation is not given but learned
  - ➢ We can call the distribution $p_\varphi(x)$ to make the dependency on φ clear

- Learning: Given a dataset $\mathcal{D} = \{\boldsymbol{x}^{(j)}\}_{j=1}^{N}$ (usually consisting of i.i.d. samples), find the parameters φ that best explain the data
  - Typically done by maximizing the marginal log-likelihood

$$\max_{\varphi} \log p_\varphi(\mathcal{D}) = \max_{\varphi} \frac{1}{N} \sum_{\boldsymbol{x}^{(j)} \in \mathcal{D}} \log p_\varphi(\boldsymbol{x}^{(j)})$$

Learning with Normalizing Flows

- **Reverse parametrization (density estimation):**



*TUM Lab Course:
- Lukas Rinder
- Markus Kittel
- Murat Can

Forward Parametrization

- We parametrize the transformation f that we know analytically.
  - $f_\theta(\boldsymbol{z}) = \boldsymbol{x}$ is computable and parameter θ can be learned

- We know that the inverse f^{-1} exists, but we might not know it analytically.
  - $f_\theta^{-1}(\boldsymbol{x}) = \boldsymbol{z}$ might not be (easily) computable

- The change of variables formula with forward parametrization is

$$p_2(\boldsymbol{x}) = p_1(\boldsymbol{z}) \cdot \left|\det\left(\frac{\partial f_\theta(\boldsymbol{z})}{\partial \boldsymbol{z}}\right)\right|^{-1}$$

  - The formula only uses the known parametrized function f_θ.
  - Given a sample $\boldsymbol{z}^{(j)}$, we can compute a sample $\boldsymbol{x}^{(j)} \sim p_2(\boldsymbol{x})$ and $p_2(\boldsymbol{x}^{(j)})$.

Forward Parametrization

- We can also stack transformations with forward parametrization i.e.

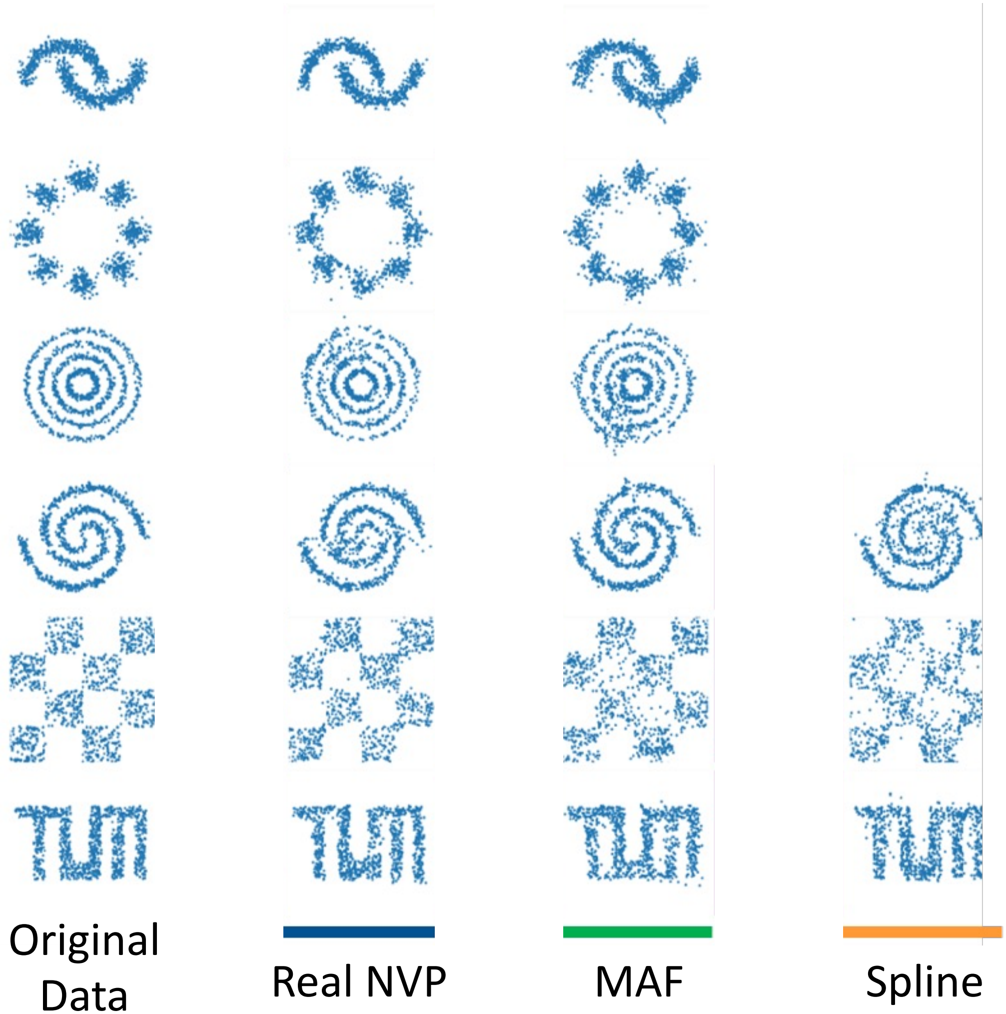
$$f_\theta = f_{\theta_K} \circ \cdots \circ f_{\theta_1}$$

- The forward parametrization enables **sampling** from the distribution $p_K(\boldsymbol{x})$.
  1. Sample $\boldsymbol{z}_{\mathbf{0}}^{(j)} \sim p_0(\boldsymbol{z}_{\mathbf{0}})$ (e.g. Gaussian or Uniform)
  2. Compute the transformations $\boldsymbol{z}_i^{(j)} = f_{\theta_i}(\boldsymbol{z}_{i-1}^{(j)})$ and $\left|\det\left(\frac{\partial f_{\theta_i}\left(\boldsymbol{z}_{i-1}^{(j)}\right)}{\partial \boldsymbol{z}_{i-1}^{(j)}}\right)\right|^{-1}$
  3. For the particular sample $\boldsymbol{x}^{(j)} = \boldsymbol{z}_K^{(j)}$, we can compute $p_K(\boldsymbol{x}^{(j)})$

- Forward pointer: This is exactly what we need in Variational Inference
  - Sample $\boldsymbol{x}$ from a distribution q and compute the probability $q(\boldsymbol{x})$ for this sample

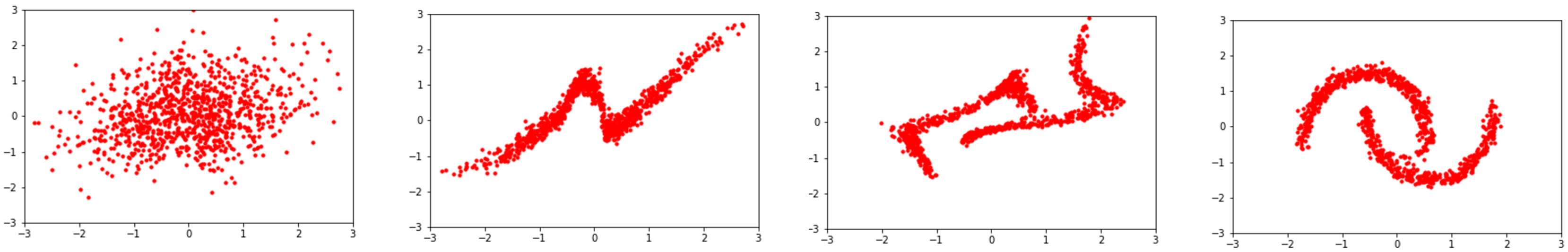
Forward Parametrization

Forward parametrization (Sampling):



*TUM Lab Course:
- Lukas Rinder
- Markus Kittel
- Murat Can

Transformation via Stacking



Visualization of a 4 layer RealNVP flow

[Eric Jang blog]

Example: Generating Images

[Kingma, Dhariwal; Glow: Generative Flow with Invertible 1×1 Convolutions]

Questions – NF2

1. For which $\boldsymbol{x}$ is it possible to compute $p(\boldsymbol{x})$ with the forward parametrization?

2. Suppose the forward transformation is defined as $f(z) = \exp(-z^n)$.
   What is the corresponding inverse transformation $g(x)$?
   Does $g(x)$ exist for any n?

3. Suppose the forward transformation is defined as the sigmoid function.
   What is the corresponding inverse transformation?

Roadmap

- Deep Generative Models
  1. Introduction
  2. **Normalizing Flows**
     - Change of Variables Formula
     - Forward and Reverse Parametrization
     - **Jacobian Determinant Computation**
  3. Variational Inference
  4. Variational Autoencoder
  5. Generative Adversarial Networks
  6. Denoising Diffusion

Jacobian Determinant Computation

- The change of variables formula involves the Jacobian determinant

$$p_2(\boldsymbol{x}) = p_1(f^{-1}(\boldsymbol{x})) \cdot \left|\det\left(\frac{\partial f^{-1}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\right|$$

- Jacobian computation can be hard/slow:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix}; \qquad g(\boldsymbol{x}) = \begin{bmatrix} g_1(\boldsymbol{x}) \\ \vdots \\ g_D(\boldsymbol{x}) \end{bmatrix}; \qquad J_g = \begin{bmatrix} \frac{\partial g_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial g_1(\boldsymbol{x})}{\partial x_D} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_D(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial g_D(\boldsymbol{x})}{\partial x_D} \end{bmatrix}$$

- How to compute effectively the Jacobian determinant ?
  - **Diagonal Jacobian**
  - **Triangular Jacobian**
  - **Full Jacobian**

Determinant properties

- Determinant of inverse: $$\det(A^{-1}) = \frac{1}{\det(A)}$$

- Determinant and eigenvalues: $$\det(A) = \prod_{i=1}^{D} \lambda_i$$

  $$eigenvalues(A) = \{\lambda_i; \mathrm{i} = 1..\mathrm{D}\}$$

- Determinant and block matrices: $$\det(A) = \det(B)\det(C)$$

  $$A = \begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix}$$

Diagonal Jacobian

- The function is applied element wise i.e.

$$g(\boldsymbol{x}) = \begin{bmatrix} g_1(x_1) \\ \vdots \\ g_D(x_D) \end{bmatrix}$$

- The Jacobian is a diagonal matrix i.e.

$$J_g = \begin{bmatrix} \frac{\partial g_1(x_1)}{\partial x_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial g_D(x_D)}{\partial x_D} \end{bmatrix}$$

- The determinant is the product of the diagonal elements ($O(D)$ complexity) i.e.

$$\det(J_g) = \prod_{i=1}^{D} \frac{\partial g_i(x_i)}{\partial x_i}$$

Triangular Jacobian

- The function is applied as

$$g(\boldsymbol{x}) = \begin{bmatrix} g_1(x_1) \\ \vdots \\ g_D(x_1, \dots, x_D) \end{bmatrix}$$

- The Jacobian is a triangular matrix i.e.

$$J_g = \begin{bmatrix} \frac{\partial g_1(x_1)}{\partial x_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ \frac{\partial g_D(x_1, \dots, x_D)}{\partial x_1} & \cdots & \frac{\partial g_D(x_1, \dots, x_D)}{\partial x_D} \end{bmatrix}$$

- The determinant is the product of the diagonal elements ($O(D)$ complexity) i.e.

$$\det(J_g) = \prod_{i=1}^{D} \frac{\partial g_i(\boldsymbol{x})}{\partial x_i}$$

- Examples:
  - Autoregressive flows

Full Jacobian

- The function is applied in the most general form

$$g(\boldsymbol{x}) = \begin{bmatrix} g_1(\boldsymbol{x}) \\ \vdots \\ g_D(\boldsymbol{x}) \end{bmatrix}$$

- The Jacobian is

$$J_g = \begin{bmatrix} \frac{\partial g_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial g_1(\boldsymbol{x})}{\partial x_D} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_D(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial g_D(\boldsymbol{x})}{\partial x_D} \end{bmatrix}$$

- The determinant can be computed with LU decomposition in $O(D^3)$ complexity
  - $J_g = LU$ where L lower triangular matrix and U upper triangular matrix.
  - $\det(J_g) = \det(L)\det(U)$ where $\det(L)$ and $\det(U)$ are diagonal products.
- Alternative for full Jacobian:
  - Continuous-time flows

Questions – NF3

1. Let's assume you get the following Jacobian:

   $$\begin{bmatrix} \frac{\partial g_1(x_1)}{\partial x_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ \frac{\partial g_D(x_1, \dots, x_D)}{\partial x_1} & \cdots & 0 \end{bmatrix}$$

   How expensive is it to compute the determinant?
   Can you comment on this in the context of NFs?

2. What is the complexity to compute the Jacobian determinant of an arbitrary valid transformation?

- Considering high-dimensional data (i.e. D is high), what type of Jacobian would you use?

References

- [1] Eric Jang blog : https://blog.evjang.com/2018/01/nf1.html
- [2] Lilian Weng blog : https://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html
- [3] Normalizing Flows for Probabilistic Modeling and Inference: https://arxiv.org/abs/1912.02762

External Sources

Web tutorial:

- Adam Kosiorek: http://akosiorek.github.io/ml/2018/04/03/norm_flows.html
- Eric Jang blog : https://blog.evjang.com/2018/01/nf1.html
- CS236 - Fall 2019 (Stanford) : https://deepgenerativemodels.github.io/notes/flow/

Survey papers:

- Normalizing Flows: An Introduction and Review of Current Methods: https://arxiv.org/pdf/1908.09257.pdf
- Normalizing Flows for Probabilistic Modeling and Inference: https://arxiv.org/abs/1912.02762

Video:

- What are normalizing flows ?: https://www.youtube.com/watch?v=i7LjDvsLWCg

Advances Machine Learning:
Deep Generative Models

Normalizing Flows

lecturer: Prof. Dr. Stephan Günnemann
www.daml.in.tum.de

Summer Term 23

Data Analytics and
Machine Learning