Variational Mixture of Normalizing Flows

Guilherme Grijó Pen Freitas Pires

November 15, 2019

Thesis to obtain the Master of Science degree in
Electrical and Computer Engineering

TÉCNICO LISBOA

Table of Contents

TÉCNICO LISBOA

Introduction and Motivation

- Deep generative models have been an active area of research, with promising results.

Introduction and Motivation

- Deep generative models have been an active area of research, with promising results.
  - **Implicit distributions**: Generative adversarial networks [Goodfellow et al., 2014], Variational Autoencoder [Kingma and Welling, 2014]
    - Don't allow explicit access to the density function

Introduction and Motivation

- Deep generative models have been an active area of research, with promising results.
  - **Implicit distributions**: Generative adversarial networks [Goodfellow et al., 2014], Variational Autoencoder [Kingma and Welling, 2014]
    - Don't allow explicit access to the density function
  - **Explicit distributions**: Normalizing flows [Rezende and Mohamed, 2015]
    - Allow explicit access to the density function
    - Lack an approach to introduce discrete structure (multi-modality) in the modelled distribution.

Introduction and Motivation: Goal

– The goal of this work was the development of a mixture of flexible distributions.

TÉCNICO LISBOA

Introduction and Motivation: Goal

- The goal of this work was the development of a mixture of flexible distributions.
- This requires answering two questions:

Introduction and Motivation: Goal

- The goal of this work was the development of a mixture of flexible distributions.
- This requires answering two questions:
  - What should be the “family” of the mixture components?

Introduction and Motivation: Goal

- The goal of this work was the development of a mixture of flexible distributions.
- This requires answering two questions:
  - What should be the “family” of the mixture components?
  - How should the mixture components’ parameters be estimated?

Outline

- Mixture Models

Outline

- Mixture Models
- Variational Inference

Outline

- Mixture Models
- Variational Inference
  - The chosen framework for estimating the parameters of the proposed model

Outline

- Mixture Models
- Variational Inference
  - The chosen framework for estimating the parameters of the proposed model
- Normalizing Flows

Outline

- Mixture Models
- Variational Inference
  - The chosen framework for estimating the parameters of the proposed model
- Normalizing Flows
  - The chosen family for the mixture model components

Outline

- Mixture Models
- Variational Inference
  - The chosen framework for estimating the parameters of the proposed model
- Normalizing Flows
  - The chosen family for the mixture model components
- Variational Mixture of Normalizing Flows

Outline

- Mixture Models
- Variational Inference
  - The chosen framework for estimating the parameters of the proposed model
- Normalizing Flows
  - The chosen family for the mixture model components
- Variational Mixture of Normalizing Flows
- Experiments and results

Outline

- Mixture Models
- Variational Inference
  - The chosen framework for estimating the parameters of the proposed model
- Normalizing Flows
  - The chosen family for the mixture model components
- Variational Mixture of Normalizing Flows
- Experiments and results
- Conclusions and future work

Table of Contents

Mixture Models: Definition

- A mixture model is used to model data that is assumed to contain subgroups.

Mixture Models: Definition

- A mixture model is used to model data that is assumed to contain subgroups.
- Typically, it is assumed that the “subgroup-conditional” distributions belong to the same family, but have different parameters.

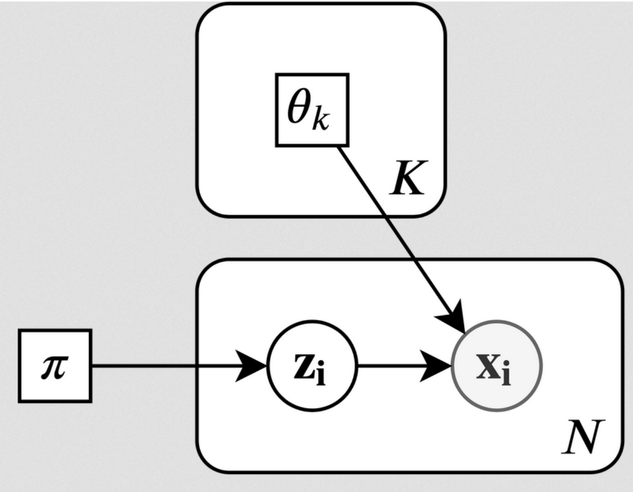
Mixture Models: Definition

- A mixture model is used to model data that is assumed to contain subgroups.
- Typically, it is assumed that the "subgroup-conditional" distributions belong to the same family, but have different parameters.
- Formally, a mixture model's joint distribution (for a single instance $\boldsymbol{x}$) is given by:

  $$p(\boldsymbol{x}, c) = p(\boldsymbol{x}|c)p(c),$$

  where c is the latent variable that indexes the subgroup to which $\boldsymbol{x}$ belongs

Mixture Models: Plate diagram



Mixture Models: Mixture of Gaussians

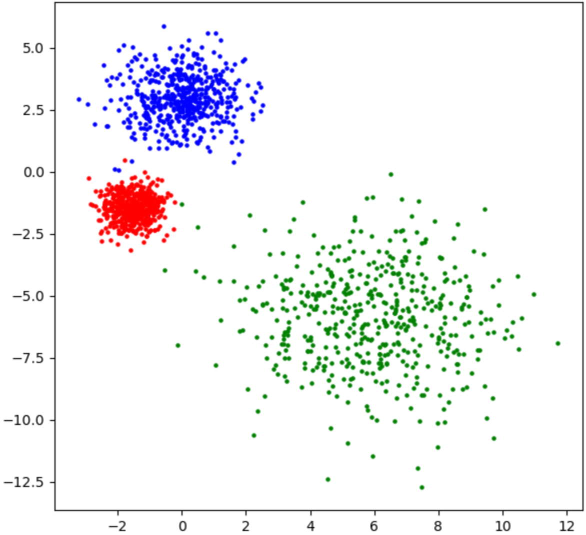


Table of Contents

Normalizing Flows: Change of Variables

Given

$$\begin{cases} \boldsymbol{z} \sim p(\boldsymbol{z}) \\ \boldsymbol{x} = g(\boldsymbol{z}; \boldsymbol{\theta}) \end{cases}$$

Normalizing Flows: Change of Variables

Given

$$\begin{cases} \boldsymbol{z} \sim p(\boldsymbol{z}) \\ \boldsymbol{x} = g(\boldsymbol{z}; \boldsymbol{\theta}) \end{cases}$$

then:

$$\begin{aligned} f_X(\boldsymbol{x}) &= f_Z(g^{-1}(\boldsymbol{x};\boldsymbol{\theta}))\left|\det\left(\frac{d}{d\boldsymbol{x}} g^{-1}(\boldsymbol{x};\boldsymbol{\theta})\right)\right| \\ &= f_Z(g^{-1}(\boldsymbol{x};\boldsymbol{\theta}))\left|\det\left(\frac{d}{d\boldsymbol{z}} g(\boldsymbol{z};\boldsymbol{\theta})\Big|_{\boldsymbol{z}=g^{-1}(\boldsymbol{x};\boldsymbol{\theta})}\right)\right|^{-1} \end{aligned}$$

Normalizing Flows: Change of Variables

Given

$$\begin{cases} \boldsymbol{z} \sim p(\boldsymbol{z}) \\ \boldsymbol{x} = g(\boldsymbol{z}; \boldsymbol{\theta}) \end{cases}$$

then:

$$\begin{aligned} f_X(\boldsymbol{x}) &= f_Z(g^{-1}(\boldsymbol{x}; \boldsymbol{\theta})) \left| \det \left(\frac{d}{d\boldsymbol{x}} g^{-1}(\boldsymbol{x}; \boldsymbol{\theta}) \right) \right| \\ &= f_Z(g^{-1}(\boldsymbol{x}; \boldsymbol{\theta})) \left| \det \left(\frac{d}{d\boldsymbol{z}} g(\boldsymbol{z}; \boldsymbol{\theta}) \Big|_{\boldsymbol{z}=g^{-1}(\boldsymbol{x}; \boldsymbol{\theta})} \right) \right|^{-1} \end{aligned}$$

This can be optimized w.r.t. $\boldsymbol{\theta}$, so as to approximate an arbitrary distribution

Normalizing Flows: Change of Variables

The described in the previous slide can be useful if

Normalizing Flows: Change of Variables

The described in the previous slide can be useful if

- The base density has a closed form expression and is easy to sample from

TÉCNICO LISBOA

Normalizing Flows: Change of Variables

The described in the previous slide can be useful if

- The base density has a closed form expression and is easy to sample from
- The determinant of the Jacobian of g is computationally cheap - not the case, in general

Normalizing Flows: Change of Variables

The described in the previous slide can be useful if

- The base density has a closed form expression and is easy to sample from
- The determinant of the Jacobian of g is computationally cheap - not the case, in general
- The gradient of the determinant of the Jacobian of g w.r.t $\boldsymbol{\theta}$ is computationally cheap

Normalizing Flows: Change of Variables

The framework of Normalizing Flows consists of composing several transformations that fulfill the three listed conditions.

Normalizing Flows: Change of Variables

The framework of Normalizing Flows consists of composing several transformations that fulfill the three listed conditions.

I.e., the function g is a composition of L functions h_ℓ, $\ell = 0, 1, ..., L-1$.

Normalizing Flows: Change of Variables

The framework of Normalizing Flows consists of composing several transformations that fulfill the three listed conditions.

I.e., the function g is a composition of L functions h_ℓ, $\ell = 0, 1, ..., L-1$. Applying the formula to g, and taking the logarithm, yields:

$$\log f_X(\boldsymbol{x}) = \log f_Z(g^{-1}(\boldsymbol{x})) - \sum_{\ell=0}^{L-1} \log \left| \det \left(\frac{d}{d\boldsymbol{x}_{\boldsymbol{\ell}}} h_\ell(\boldsymbol{x}_{\boldsymbol{\ell}}) \right) \right|.$$

Normalizing Flows: Affine Coupling Layer

An example of such a transformation is the Affine Coupling Layer {[Dinh, Sohl-Dickstein, and Bengio, 2017]}.

Splitting $\boldsymbol{z}$ into $(\boldsymbol{z_1}, \boldsymbol{z_2})$,

Normalizing Flows: Affine Coupling Layer

An example of such a transformation is the Affine Coupling Layer {[Dinh, Sohl-Dickstein, and Bengio, 2017]}.

Splitting $\boldsymbol{z}$ into $(\boldsymbol{z_1}, \boldsymbol{z_2})$,

$$\begin{cases} \boldsymbol{x_1} & = \boldsymbol{z_1} \odot \exp\left(s(\boldsymbol{z_2})\right) + t(\boldsymbol{z_2}) \\ \boldsymbol{x_2} & = \boldsymbol{z_2}. \end{cases}$$

Normalizing Flows: Affine Coupling Layer

An example of such a transformation is the Affine Coupling Layer {[Dinh, Sohl-Dickstein, and Bengio, 2017]}.

Splitting z into (z_1, z_2),

$$\begin{cases} x_1 & = z_1 \odot \exp\big(s(z_2)\big) + t(z_2) \\ x_2 & = z_2 . \end{cases}$$

This transformation has the following Jacobian matrix:

$$J_{f(z)} = \begin{bmatrix} \frac{\partial x_1}{\partial z_1} & \frac{\partial x_1}{\partial z_2} \\ \frac{\partial x_2}{\partial z_1} & \frac{\partial x_2}{\partial z_2} \end{bmatrix} = \begin{bmatrix} \text{diag}\Big(\exp\big(s(z_2)\big)\Big) & \frac{\partial x_1}{\partial z_2} \\ \mathbf{0} & I \end{bmatrix}$$

Table of Contents

TÉCNICO LISBOA

Variational Inference: Preamble

Consider a joint probability distribution $p(\boldsymbol{x}, \boldsymbol{z})$.

Variational Inference: Preamble

Consider a joint probability distribution $p(\boldsymbol{x}, \boldsymbol{z})$. Suppose $\boldsymbol{x}$ is observed and $\boldsymbol{z}$ is latent.

Variational Inference: Preamble

Consider a joint probability distribution $p(\boldsymbol{x}, \boldsymbol{z})$. Suppose $\boldsymbol{x}$ is observed and $\boldsymbol{z}$ is latent. If we want to infer the most probable values of $\boldsymbol{z}$, given $\boldsymbol{x}$, by Bayes' Law:

$$\begin{aligned} p(\boldsymbol{z}|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{x})} \\ &= \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{\int p(\boldsymbol{x}|\boldsymbol{z}')p(\boldsymbol{z}')d\boldsymbol{z}'} \end{aligned}$$

Variational Inference: Preamble

Consider a joint probability distribution $p(\boldsymbol{x}, \boldsymbol{z})$. Suppose $\boldsymbol{x}$ is observed and $\boldsymbol{z}$ is latent. If we want to infer the most probable values of $\boldsymbol{z}$, given $\boldsymbol{x}$, by Bayes' Law:

$$
\begin{aligned}
p(\boldsymbol{z}|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{x})} \\
&= \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{\int p(\boldsymbol{x}|\boldsymbol{z}')p(\boldsymbol{z}')d\boldsymbol{z}'}
\end{aligned}
$$

Problem: The integral in the denominator is intractable for most interesting models.

Variational Inference: Preamble

Consider a joint probability distribution $p(\boldsymbol{x}, \boldsymbol{z})$. Suppose $\boldsymbol{x}$ is observed and $\boldsymbol{z}$ is latent. If we want to infer the most probable values of $\boldsymbol{z}$, given $\boldsymbol{x}$, by Bayes' Law:

$$\begin{aligned} p(\boldsymbol{z}|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{x})} \\ &= \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{\int p(\boldsymbol{x}|\boldsymbol{z'})p(\boldsymbol{z'})d\boldsymbol{z'}} \end{aligned}$$

Problem: The integral in the denominator is intractable for most interesting models.

- Variational inference is an approximate inference framework, that can be used to overcome this intractability.

Variational Inference: Goal

Given a parametric family of distributions $q(\boldsymbol{z};\boldsymbol{\lambda})$, find the parameters $\boldsymbol{\lambda}$ that minimize the Kullback-Leibler divergence between $q(\boldsymbol{z};\boldsymbol{\lambda})$ and $p(\boldsymbol{z}|\boldsymbol{x})$

Variational Inference: ELBO

$$KL(q||p) = \int q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})} d\boldsymbol{z}$$

Variational Inference: ELBO

$$
\begin{aligned}
KL(q||p) &= \int q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})} d\boldsymbol{z} \\
&= \int q(\boldsymbol{z})(\log q(\boldsymbol{z}) - \log p(\boldsymbol{z}|\boldsymbol{x}))d\boldsymbol{z}
\end{aligned}
$$

Variational Inference: ELBO

$$
\begin{aligned}
KL(q||p) &= \int q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})} d\boldsymbol{z} \\
&= \int q(\boldsymbol{z})(\log q(\boldsymbol{z}) - \log p(\boldsymbol{z}|\boldsymbol{x}))d\boldsymbol{z} \\
&= \int q(\boldsymbol{z})(\log q(\boldsymbol{z}) - (\log p(\boldsymbol{x}, \boldsymbol{z}) - \log p(\boldsymbol{x})))d\boldsymbol{z}
\end{aligned}
$$

Variational Inference: ELBO

$$
\begin{aligned}
KL(q||p) &= \int q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})} d\boldsymbol{z} \\
&= \int q(\boldsymbol{z})(\log q(\boldsymbol{z}) - \log p(\boldsymbol{z}|\boldsymbol{x}))d\boldsymbol{z} \\
&= \int q(\boldsymbol{z})(\log q(\boldsymbol{z}) - (\log p(\boldsymbol{x}, \boldsymbol{z}) - \log p(\boldsymbol{x})))d\boldsymbol{z} \\
&= \mathbb{E}_q[\log q(\boldsymbol{z})] - \mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{z})] + \log p(\boldsymbol{x})
\end{aligned}
$$

Variational Inference: ELBO

$$
\begin{aligned}
KL(q||p) &= \int q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})} d\boldsymbol{z} \\
&= \int q(\boldsymbol{z})(\log q(\boldsymbol{z}) - \log p(\boldsymbol{z}|\boldsymbol{x}))d\boldsymbol{z} \\
&= \int q(\boldsymbol{z})(\log q(\boldsymbol{z}) - (\log p(\boldsymbol{x}, \boldsymbol{z}) - \log p(\boldsymbol{x})))d\boldsymbol{z} \\
&= \mathbb{E}_q[\log q(\boldsymbol{z})] - \mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{z})] + \log p(\boldsymbol{x})
\end{aligned}
$$

Which yields the lower bound (ELBO):

$$
\begin{aligned}
ELBO(q) &= \mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{z})] - \mathbb{E}_q[\log q(\boldsymbol{z})] \\
&= \mathbb{E}_q[\log p(\boldsymbol{x}|\boldsymbol{z})] + \mathbb{E}_q[\log p(\boldsymbol{z})] - \mathbb{E}_q[\log q(\boldsymbol{z})]
\end{aligned}
$$

Table of Contents

VMoNF: Introduction

Is it possible to combine the ideas from the previous sections, to obtain a mixture of flexible models?

VMoNF: Definition

Recall the ELBO:

$$ELBO(q) = \mathbb{E}_q[\log p(\boldsymbol{x}|\boldsymbol{z})] + \mathbb{E}_q[\log p(\boldsymbol{z})] - \mathbb{E}_q[\log q(\boldsymbol{z})]$$

TÉCNICO LISBOA

VMoNF: Definition

Recall the ELBO:

$$ELBO(q) = \mathbb{E}_q[\log p(\boldsymbol{x}|\boldsymbol{z})] + \mathbb{E}_q[\log p(\boldsymbol{z})] - \mathbb{E}_q[\log q(\boldsymbol{z})]$$

Let the variational posterior $q(z|x)$ be parameterized by a neural network.

VMoNF: Definition

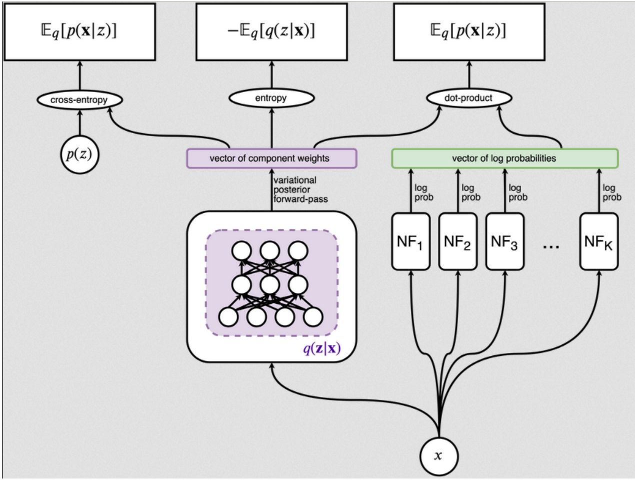
Recall the ELBO:

$$ELBO(q) = \mathbb{E}_q[\log p(\boldsymbol{x}|\boldsymbol{z})] + \mathbb{E}_q[\log p(\boldsymbol{z})] - \mathbb{E}_q[\log q(\boldsymbol{z})]$$

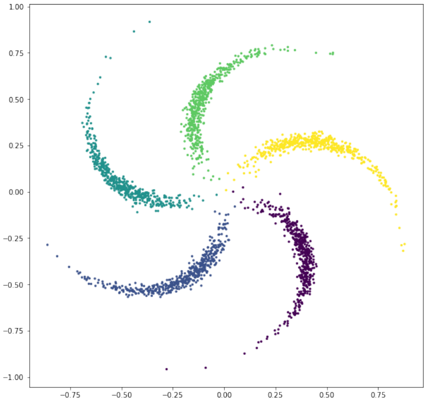
Let the variational posterior $q(z|x)$ be parameterized by a neural network.

We optimize this objective, by **jointly** learning the variational posterior and the generative components.

VMoNF: Overview

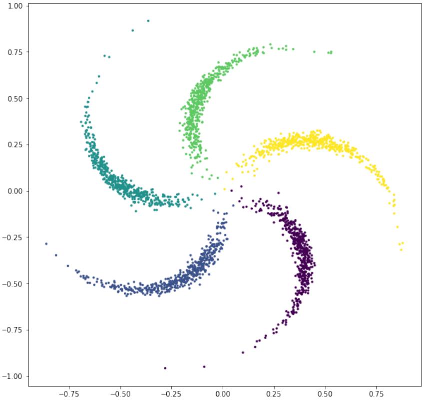


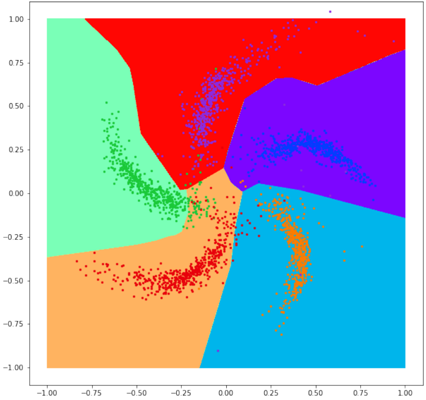
VMoNF: Experiments - Pinwheel (5 wings)



TÉCNICO LISBOA

VMoNF: Experiments - Pinwheel (5 wings)



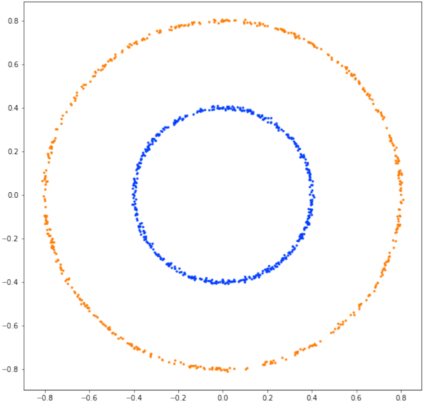


TÉCNICO LISBOA

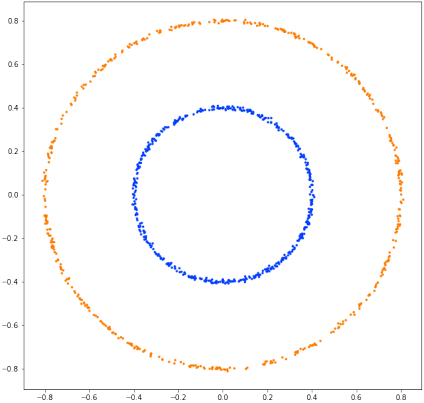
VMoNF: Experiments - Pinwheel (3 wings)

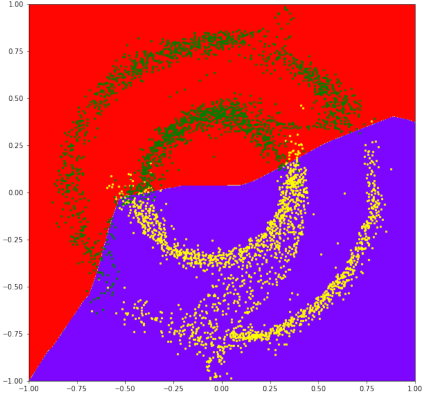
Trainining Animation

VMoNF: Experiments - 2 Circles

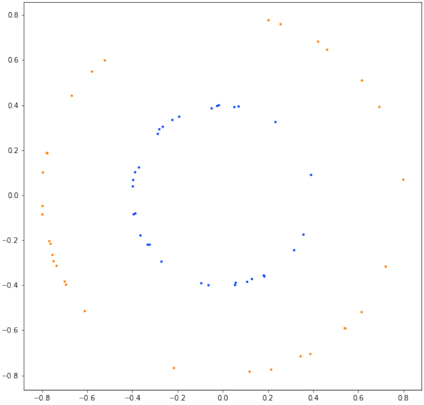


VMoNF: Experiments - 2 Circles

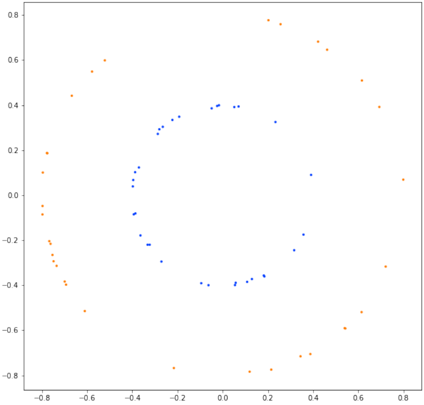


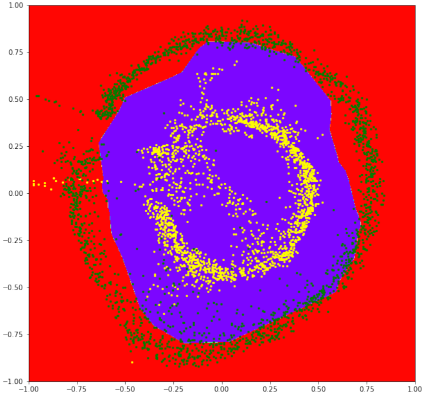


VMoNF: Experiments - 2 Circles (semi supervised)



VMoNF: Experiments - 2 Circles (semi supervised)





Note: 32 labeled points, 1024 unlabeled points

VMoNF: Experiments - MNIST

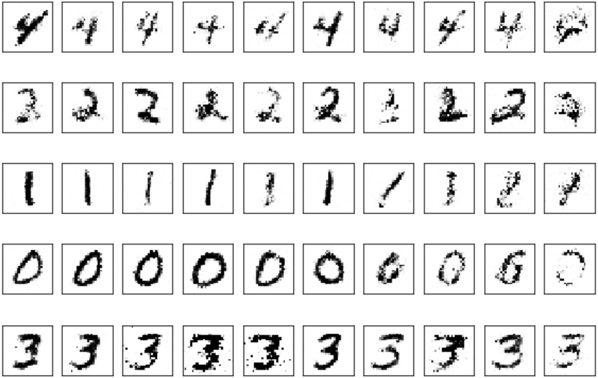


Table of Contents

Conclusions and Future Work

- Similar work is being pursued and published in prominent venues: [Dinh, Sohl-Dickstein, et al., 2019; Izmailov et al., 2019]

Conclusions and Future Work

- Similar work is being pursued and published in prominent venues: [Dinh, Sohl-Dickstein, et al., 2019; Izmailov et al., 2019]
- Formally describe the reasons why the model fails in cases like the 2 circles $\rightarrow$ Topology

Conclusions and Future Work

- Similar work is being pursued and published in prominent venues: [Dinh, Sohl-Dickstein, et al., 2019; Izmailov et al., 2019]
- Formally describe the reasons why the model fails in cases like the 2 circles $\rightarrow$ Topology
  - Investigate the effect of a consistency loss regularization term

Conclusions and Future Work

- Similar work is being pursued and published in prominent venues: [Dinh, Sohl-Dickstein, et al., 2019; Izmailov et al., 2019]
- Formally describe the reasons why the model fails in cases like the 2 circles $\rightarrow$ Topology
  - Investigate the effect of a consistency loss regularization term
- Weight-sharing between components

TÉCNICO LISBOA

Conclusions and Future Work

- Similar work is being pursued and published in prominent venues: [Dinh, Sohl-Dickstein, et al., 2019; Izmailov et al., 2019]
- Formally describe the reasons why the model fails in cases like the 2 circles $\rightarrow$ Topology
  - Investigate the effect of a consistency loss regularization term
- Weight-sharing between components
- Balance between complexities

TÉCNICO LISBOA

Conclusions and Future Work

- Similar work is being pursued and published in prominent venues: [Dinh, Sohl-Dickstein, et al., 2019; Izmailov et al., 2019]
- Formally describe the reasons why the model fails in cases like the 2 circles $\rightarrow$ Topology
  - Investigate the effect of a consistency loss regularization term
- Weight-sharing between components
- Balance between complexities
- (Controlled) component anihilation

TÉCNICO LISBOA

Thank you!