Variational Mixture of Normalizing Flows

Guilherme Grijó Pen Freitas Pires

November 20, 2019

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

Supervisor: Prof. Mário A. T. Figueiredo

TÉCNICO LISBOA

Table of Contents

TÉCNICO LISBOA

Introduction and Motivation

- Deep generative models: an active area of research

Introduction and Motivation

- Deep generative models: an active area of research
  - **Implicit distributions**: Generative adversarial networks [Goodfellow et al., 2014], Variational Autoencoder [Kingma and Welling, 2014]
    - No explicit access to the density function

Introduction and Motivation

- Deep generative models: an active area of research
  - **Implicit distributions**: Generative adversarial networks [Goodfellow et al., 2014], Variational Autoencoder [Kingma and Welling, 2014]
    - No explicit access to the density function
  - **Explicit distributions**: Normalizing flows [Rezende and Mohamed, 2015]
    - Explicit access to the density function
    - No approach to introduce discrete structure (multi-modality)

TÉCNICO LISBOA

Introduction and Motivation: Goal

- Goal of this work: **mixture of flexible distributions**.

TÉCNICO LISBOA

Introduction and Motivation: Goal

- Goal of this work: **mixture of flexible distributions**.
- Two questions:

Introduction and Motivation: Goal

- Goal of this work: **mixture of flexible distributions**.
- Two questions:
  - What should the **mixture components** be?

Introduction and Motivation: Goal

- Goal of this work: **mixture of flexible distributions**.
- Two questions:
  - What should the **mixture components** be?
  - How should their **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

TÉCNICO LISBOA

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

TÉCNICO LISBOA

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

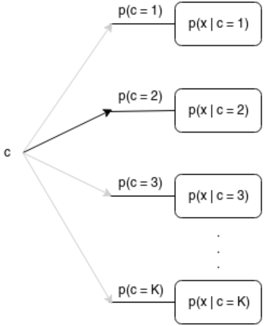
- Mixture model: used to model data that contains **subgroups**.

Mixture Models: Definition

- Mixture model: used to model data that contains **subgroups**.
- "Subgroup-conditional" distributions (typically) in the same family

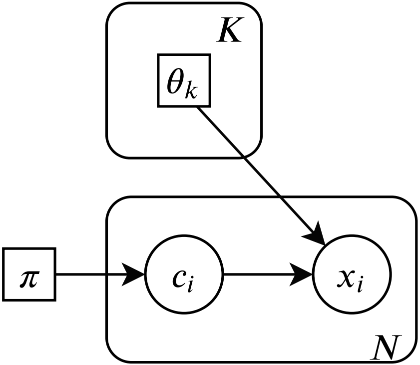
Mixture Models: Definition

- Mixture model: used to model data that contains **subgroups**.
- "Subgroup-conditional" distributions (typically) in the same family

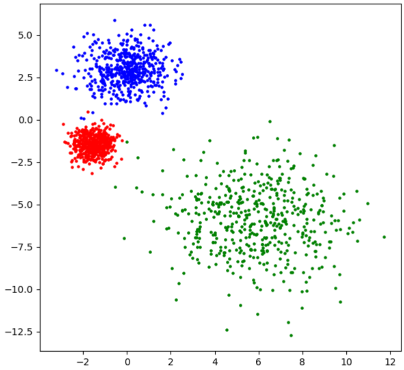


TÉCNICO LISBOA

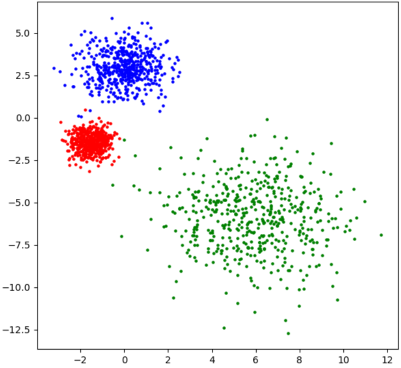
Mixture Models: Plate diagram

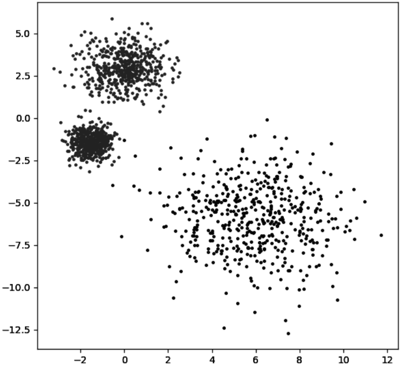


Mixture Models: Mixture of Gaussians



Mixture Models: Mixture of Gaussians





TÉCNICO LISBOA

Table of Contents

TÉCNICO LISBOA

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:

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

TÉCNICO LISBOA

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. $\boldsymbol{\theta}$, to approximate an arbitrary distribution

Requirements for feasibility

Normalizing Flows: Change of Variables

Requirements for feasibility

1. Base density: **closed form** and **easy to sample** from

Normalizing Flows: Change of Variables

Requirements for feasibility

1. Base density: **closed form** and **easy to sample** from
2. **Determinant** of the **Jacobian** of g: computationally cheap

TÉCNICO LISBOA

Normalizing Flows: Change of Variables

Requirements for feasibility

1. Base density: **closed form** and **easy to sample** from
2. **Determinant** of the **Jacobian** of g: computationally cheap
3. **Gradient** of b w.r.t $\boldsymbol{\theta}$: computationally cheap

Normalizing Flows: composition of several “good” transformations

Normalizing Flows: Change of Variables

Normalizing Flows: composition of several “good” transformations I.e., $g = h_{L-1} \odot h_{L-2} \odot ... \odot h_1 \odot h_0$

Normalizing Flows: Change of Variables

Normalizing Flows: composition of several "good" transformations I.e., $g = h_{L-1} \odot h_{L-2} \odot \ldots \odot h_1 \odot h_0$ Applying the formula to g, and taking the logarithm:

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

Normalizing Flows: Affine Coupling Layer

An example: 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: 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: Affine Coupling Layer {[Dinh, Sohl-Dickstein, and Bengio, 2017]}.

Splitting $\boldsymbol{z}$ into $(\boldsymbol{z}_1, \boldsymbol{z}_2)$,

$$\begin{cases} \boldsymbol{x}_1 & = \boldsymbol{z}_1 \odot \exp\left(s(\boldsymbol{z}_2)\right) + t(\boldsymbol{z}_2) \\ \boldsymbol{x}_2 & = \boldsymbol{z}_2. \end{cases}$$

The respective Jacobian matrix:

$$J_{f(z)} = \begin{bmatrix} \frac{\partial \boldsymbol{x}_1}{\partial \boldsymbol{z}_1} & \frac{\partial \boldsymbol{x}_1}{\partial \boldsymbol{z}_2} \\ \frac{\partial \boldsymbol{x}_2}{\partial \boldsymbol{z}_1} & \frac{\partial \boldsymbol{x}_2}{\partial \boldsymbol{z}_2} \end{bmatrix} = \begin{bmatrix} \operatorname{diag}\Big(\exp\left(s(\boldsymbol{z}_2)\right)\Big) & \frac{\partial \boldsymbol{x}_1}{\partial \boldsymbol{z}_2} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix}$$

TÉCNICO LISBOA

Table of Contents

TÉCNICO LISBOA

Variational Inference: Preamble

Joint probability distribution $p(\boldsymbol{x}, \boldsymbol{c})$.

Variational Inference: Preamble

Joint probability distribution $p(\boldsymbol{x}, \boldsymbol{c})$. $\boldsymbol{x}$ is observed and $\boldsymbol{c}$ is latent.

Variational Inference: Preamble

Joint probability distribution $p(\boldsymbol{x}, \boldsymbol{c})$. $\boldsymbol{x}$ is observed and $\boldsymbol{c}$ is latent.
Inference about $\boldsymbol{z}$, given $\boldsymbol{x}$, by Bayes' Law:

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

TÉCNICO LISBOA

Variational Inference: Preamble

Joint probability distribution $p(\boldsymbol{x}, \boldsymbol{c})$. $\boldsymbol{x}$ is observed and $\boldsymbol{c}$ is latent.
Inference about $\boldsymbol{z}$, 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 is normally intractable

Variational Inference: Preamble

Joint probability distribution $p(\boldsymbol{x}, \boldsymbol{c})$. $\boldsymbol{x}$ is observed and $\boldsymbol{c}$ is latent.
Inference about $\boldsymbol{z}$, given $\boldsymbol{x}$, by Bayes' Law:

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

Problem: The integral is normally intractable

- Variational inference: an approximate inference framework to overcome this intractability.

Variational Inference: Goal

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

$$\boldsymbol{\lambda}^* = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}}\, KL(q(\boldsymbol{z}; \boldsymbol{\lambda}) || p(\boldsymbol{z}|\boldsymbol{x}))$$

Variational Inference: ELBO

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

Variational Inference: ELBO

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

Variational Inference: ELBO

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

Variational Inference: ELBO

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

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

Which yields the lower bound (ELBO):

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

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:

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

VMoNF: Definition

Recall the ELBO:

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

$q(z|x)$ is parameterized by a neural network.

VMoNF: Definition

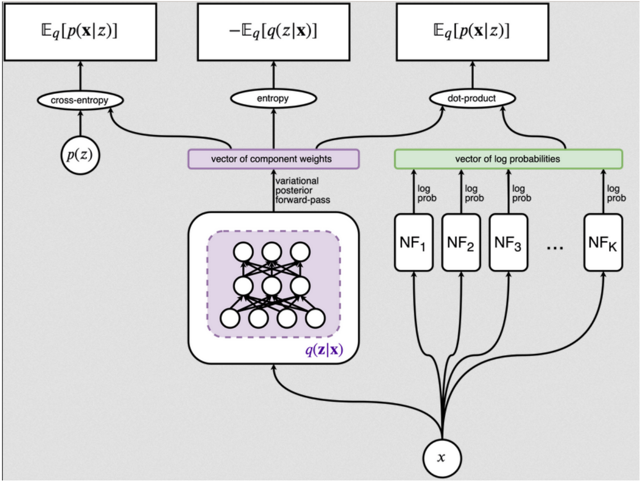
Recall the ELBO:

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

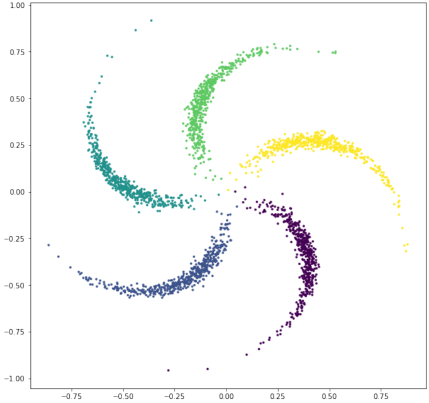
$q(z|x)$ is parameterized by a neural network.

Optimize the ELBO, by **jointly** learning the variational posterior and the generative components.

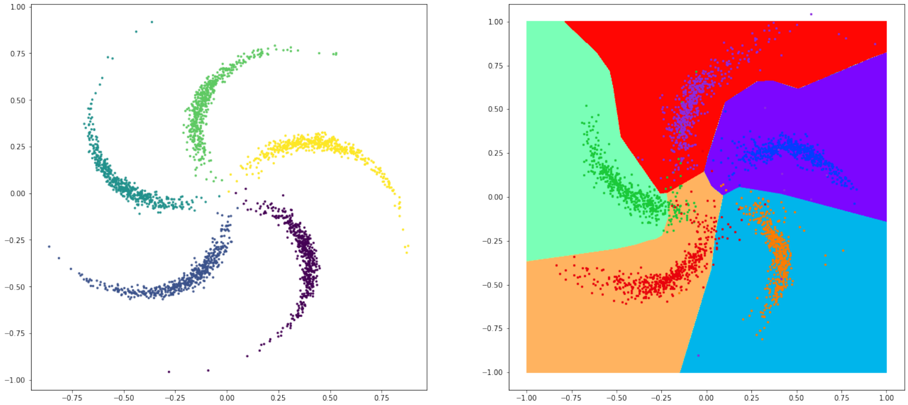
VMoNF: Overview



VMoNF: Experiments - Pinwheel (5 wings)



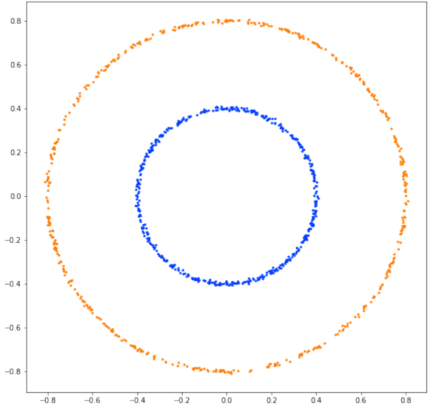
VMoNF: Experiments - Pinwheel (5 wings)



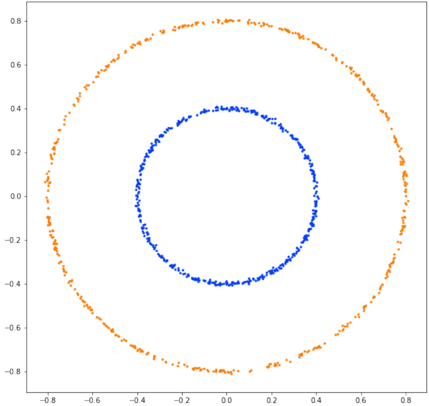
VMoNF: Experiments - Pinwheel (3 wings)

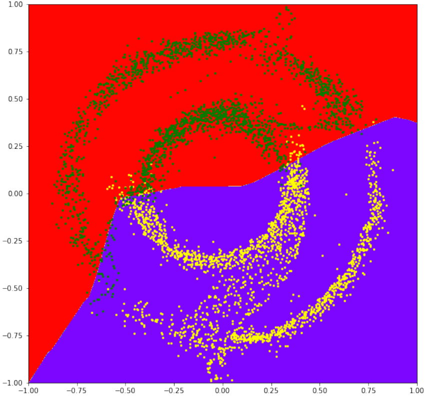
Trainining Animation

VMoNF: Experiments - 2 Circles

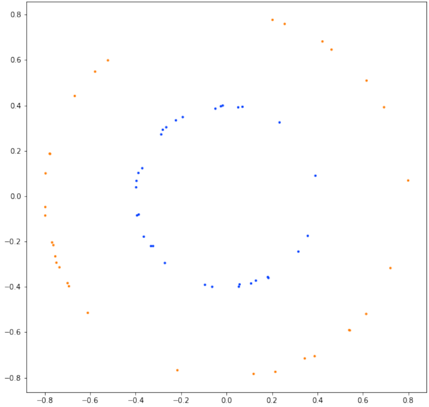


VMoNF: Experiments - 2 Circles

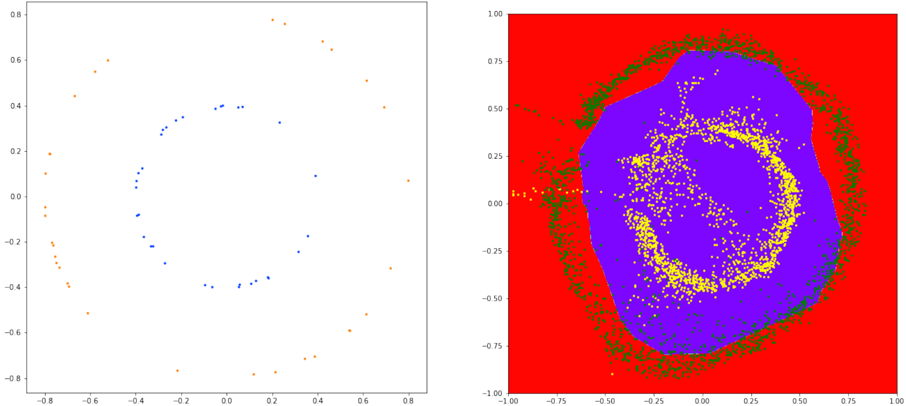




VMoNF: Experiments - 2 Circles (semi supervised)



VMoNF: Experiments - 2 Circles (semi supervised)



Note: 32 labeled points, 1024 unlabeled points

VMoNF: Experiments - MNIST

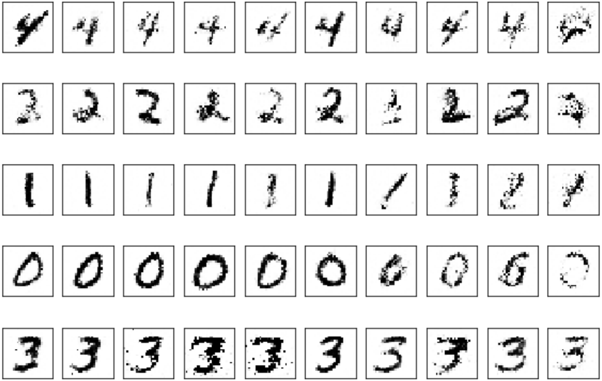


Table of Contents

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

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

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

Thank you!