Variational Mixture of Normalizing Flows

Guilherme Grijó Pen Freitas Pires

November 13, 2019

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

TÉCNICO LISBOA

Summary

This work in one sentence:

- The development of a mixture of *normalizing flows* and a (variational) training procedure for it.

Table of Contents

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

Introduction and Motivation

- This work

Introduction and Motivation

- This work
  - How to endow normalizing flows with discrete structure?

Introduction and Motivation

- This work
  - How to endow normalizing flows with discrete structure?
  - Or, how to endow mixture models with more expressiveness/flexibility?

Outline

- Introductory concepts on probabilistic modelling

TÉCNICO LISBOA

Outline

- Introductory concepts on probabilistic modelling
- Variational Inference

Outline

- Introductory concepts on probabilistic modelling
- Variational Inference
  - The chosen method for optimizing the model proposed in this work

Outline

- Introductory concepts on probabilistic modelling
- Variational Inference
  - The chosen method for optimizing the model proposed in this work
- Normalizing Flows

Outline

- Introductory concepts on probabilistic modelling
- Variational Inference
  - The chosen method for optimizing the model proposed in this work
- Normalizing Flows
  - The centerpiece of the proposed model

Outline

- Introductory concepts on probabilistic modelling
- Variational Inference
  - The chosen method for optimizing the model proposed in this work
- Normalizing Flows
  - The centerpiece of the proposed model
- Variational Mixture of Normalizing Flows

Outline

- Introductory concepts on probabilistic modelling
- Variational Inference
  - The chosen method for optimizing the model proposed in this work
- Normalizing Flows
  - The centerpiece of the proposed model
- Variational Mixture of Normalizing Flows
- Experiments and results

Outline

- Introductory concepts on probabilistic modelling
- Variational Inference
  - The chosen method for optimizing the model proposed in this work
- Normalizing Flows
  - The centerpiece of the proposed model
- Variational Mixture of Normalizing Flows
- Experiments and results
- Conclusions and future work

Table of Contents

Probabilistic Modelling: Goal

Given data, find the probability distribution (commonly referred to as the *model*) that is the closest possible approximation to the true distribution of the data.

Probabilistic Modelling: Goal

Informally, via Bayes' Law:

$$p(\text{hypothesis}|\text{data}) = \frac{p(\text{data}|\text{hypothesis})p(\text{hypothesis})}{p(\text{data})}.$$

Probabilistic Modelling: Goal

Informally, via Bayes' Law:

$$p(\text{hypothesis}|\text{data}) = \frac{p(\text{data}|\text{hypothesis})p(\text{hypothesis})}{p(\text{data})}.$$

The goal of probabilistic modelling is to find the optimal hypothesis that maximizes (some form of) this expression.

Probabilistic Modelling: Hypothesis

There are infinite candidate distributions to model the data. In practice, the scope is narrowed to a class of hypothesis.

Probabilistic Modelling: Hypothesis

There are infinite candidate distributions to model the data. In practice, the scope is narrowed to a class of hypothesis.

- Parametric families: $p(\boldsymbol{x}|\boldsymbol{\theta})$

Probabilistic Modelling: Hypothesis

There are infinite candidate distributions to model the data. In practice, the scope is narrowed to a class of hypothesis.

- Parametric families: $p(\boldsymbol{x}|\boldsymbol{\theta})$
- Particular factorizations: e.g. $p(\boldsymbol{x}) = \prod_i^N p(x_i|x_{i-1})$

Probabilistic Modelling: Hypothesis

There are infinite candidate distributions to model the data. In practice, the scope is narrowed to a class of hypothesis.

- Parametric families: $p(\boldsymbol{x}|\boldsymbol{\theta})$
- Particular factorizations: e.g. $p(\boldsymbol{x}) = \prod_i^N p(x_i|x_{i-1})$
- Latent variables: $p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}$[1]

[1] In the case of discrete latent variables, the integral becomes a sum

Probabilistic Modelling: Hypothesis

There are infinite candidate distributions to model the data. In practice, the scope is narrowed to a class of hypothesis.

- Parametric families: $p(\boldsymbol{x}|\boldsymbol{\theta})$
- Particular factorizations: e.g. $p(\boldsymbol{x}) = \prod_i^N p(x_i|x_{i-1})$
- Latent variables: $p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}$[1]
- All of the above, combined

[1] In the case of discrete latent variables, the integral becomes a sum

Probabilistic Modelling: Hypothesis

Example: Mixture Models.

Probabilistic Modelling: Hypothesis

Example: Mixture Models.

A mixture model is defined as:

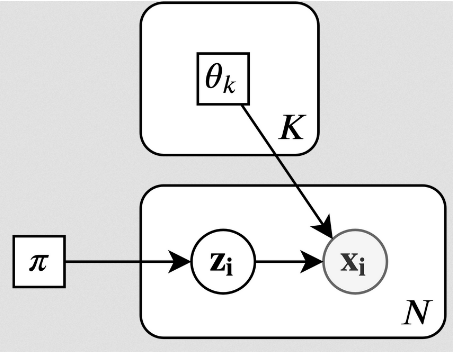
$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z})$$

Probabilistic Modelling: Hypothesis

Example: Mixture Models.

A mixture model is defined as:

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



Probabilistic Modelling: Parameter Estimation

Given data x and a parametric model $p(\boldsymbol{x}|\boldsymbol{\theta})$, to estimate $\boldsymbol{\theta}$:

Probabilistic Modelling: Parameter Estimation

Given data $\boldsymbol{x}$ and a parametric model $p(\boldsymbol{x}|\boldsymbol{\theta})$, to estimate $\boldsymbol{\theta}$:

Maximum-likelihood:

$$\begin{cases} \hat{\boldsymbol{\theta}}_{ML} = \text{argmax}_{\boldsymbol{\theta}}\, \mathcal{L}(\boldsymbol{\theta}), \\ \mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta}) \end{cases}$$

Probabilistic Modelling: Parameter Estimation

Given data x and a parametric model $p(\boldsymbol{x}|\boldsymbol{\theta})$, to estimate $\boldsymbol{\theta}$:

Maximum-likelihood:

$$\begin{cases} \hat{\boldsymbol{\theta}}_{ML} = \text{argmax}_{\boldsymbol{\theta}} \, \mathcal{L}(\boldsymbol{\theta}), \\ \mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta}) \end{cases}$$

Maximum a posteriori:

$$\begin{cases} \hat{\boldsymbol{\theta}}_{MAP} = \text{argmax}_{\boldsymbol{\theta}} \, p(\boldsymbol{\theta}|\boldsymbol{x}), \\ p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})}. \end{cases}$$

Probabilistic Modelling: Inference

Given data, $\boldsymbol{x}$, and a model $p(\boldsymbol{x}, \boldsymbol{z})$, one is generally interested in finding the posterior:

Probabilistic Modelling: Inference

Given data, $\boldsymbol{x}$, and a model $p(\boldsymbol{x}, \boldsymbol{z})$, one is generally interested in finding the posterior:

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

Probabilistic Modelling: Inference

Given data, $\boldsymbol{x}$, and a model $p(\boldsymbol{x}, \boldsymbol{z})$, one is generally interested in finding the posterior:

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

In general, the integral in the denominator is intractable.

Probabilistic Modelling: Inference

Given data, $\boldsymbol{x}$, and a model $p(\boldsymbol{x}, \boldsymbol{z})$, one is generally interested in finding the posterior:

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

In general, the integral in the denominator is intractable.

$\rightarrow$ Approximate inference is required.

Table of Contents

Variational Inference: Goal

Variational Inference is one way of dealing with the intractability previously described.

Variational Inference: Goal

Variational Inference is one way of dealing with the intractability previously described.

Given a *variational* family $q(\boldsymbol{z}; \boldsymbol{\lambda})$, find the parameters $\boldsymbol{\lambda}$ that minimze 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

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:

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

Normalizing Flows: Change of Variables

Given

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

then:

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

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

Normalizing Flows: Change of Variables

The above can be useful if

Normalizing Flows: Change of Variables

The above can be useful if

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

Normalizing Flows: Change of Variables

The above 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 above 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: 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 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 $(\boldsymbol{z_1}, \boldsymbol{z_2})$,

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

This transformation has the following Jacobian matrix:

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

Table of Contents

VMoNF: Motivation

How can we leverage the flexibility of normalizing flows, and endow it with multimodal, discrete structure, like in a mixture model?

VMoNF: Motivation

How can we leverage the flexibility of normalizing flows, and endow it with multimodal, discrete structure, like in a mixture model?

Mixture of normalizing flows.

VMoNF: Motivation

How can we leverage the flexibility of normalizing flows, and endow it with multimodal, discrete structure, like in a mixture model?

Mixture of normalizing flows. $\rightarrow$ Approximate inference is required.

VMoNF: Definition

Recall the ELBO:

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

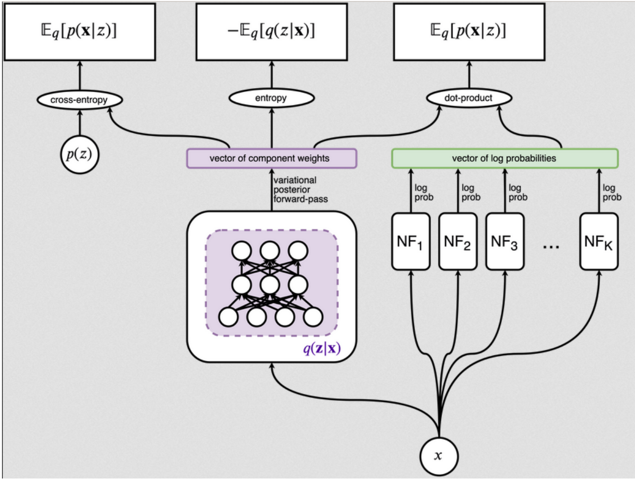
VMoNF: Definition

Recall the ELBO:

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

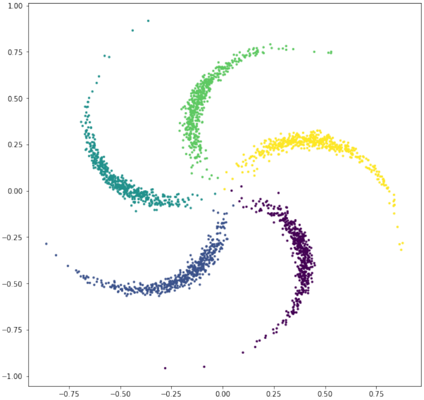
Let the variational posterior $q(z|x)$ be parameterized by a neural network. We jointly optimize this objective, hence we learn the variational posterior and the generative components simultaneously.

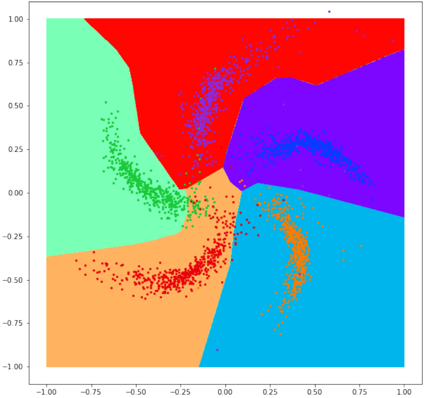
VMoNF: Overview



TÉCNICO LISBOA

VMoNF: Experiments - Pinwheel (5 wings)



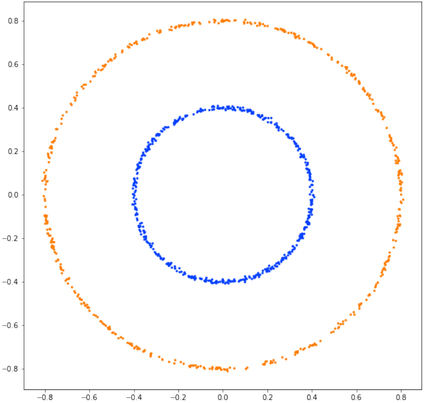


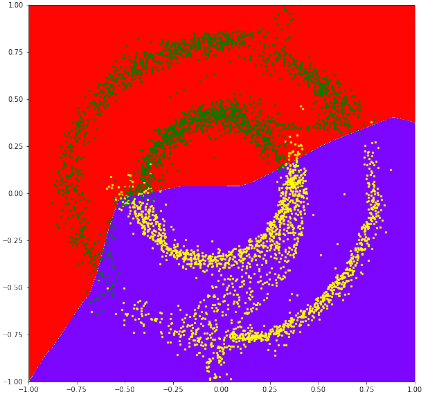
TÉCNICO LISBOA

VMoNF: Experiments - Pinwheel (3 wings)

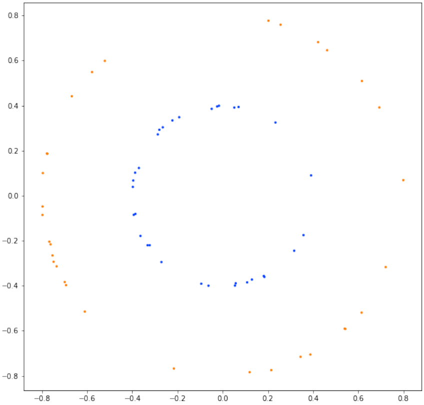
Trainining Animation

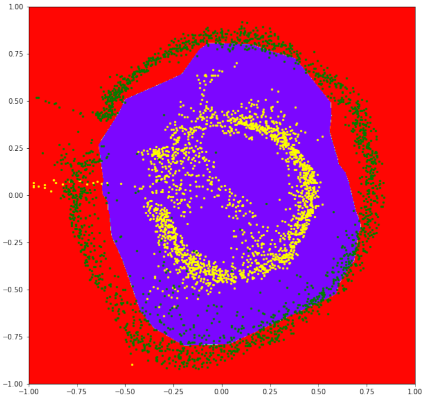
VMoNF: Experiments - 2 Circles





VMoNF: Experiments - 2 Circles (semi supervised)





Note: 32 labeled points, 1024 unlabeled points

TÉCNICO LISBOA

VMoNF: Experiments - MNIST

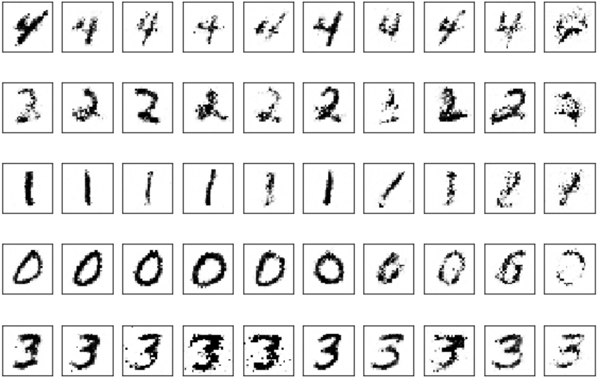


Table of Contents

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!