THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Variational Mixture of Normalizing Flows

Guilherme P. Grijó Pires

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor(s): Prof. Mário Alexandre Teles de Figueiredo

Examination Committee

Chairperson: Prof. Full Name
Supervisor: Prof. Full Name 1 (or 2)
Member of the Committee: Prof. Full Name 3

Month Year

What I cannot create, I do not understand.

- Richard Feynman

Acknowledgments

A few words about the university, financial support, research advisor, dissertation readers, faculty or other professors, lab mates, other friends and family...

Resumo

Inserir o resumo em Português aqui com o máximo de 250 palavras e acompanhado de 4 a 6 palavras-chave...

Palavras-chave: palavra-chave1, palavra-chave2,...

Abstract

Keywords:

Contents

List of Tables

List of Figures

Nomenclature

Greek symbols

α	Angle of attack.
β	Angle of side-slip.
κ	Thermal conductivity coefficient.
μ	Molecular viscosity coefficient.
ρ	Density.

Roman symbols

C_D	Coefficient of drag.
C_L	Coefficient of lift.
C_M	Coefficient of moment.
p	Pressure.
$\mathbf{u}$	Velocity vector.
u, v, w	Velocity Cartesian components.

Subscripts

∞	Free-stream condition.
i, j, k	Computational indexes.
n	Normal component.
x, y, z	Cartesian components.
ref	Reference condition.

Superscripts

*	Adjoint.
T	Transpose.

Glossary

CFD Computational Fluid Dynamics is a branch of fluid mechanics that uses numerical methods and algorithms to solve problems that involve fluid flows.

CSM Computational Structural Mechanics is a branch of structure mechanics that uses numerical methods and algorithms to perform the analysis of structures and its components.

MDO Multi-Disciplinar Optimization is an engineering technique that uses optimization methods to solve design problems incorporating two or more disciplines.

Chapter 1

Introduction

Insert your chapter material here...

1.1 Motivation

Relevance of the subject...

1.2 Topic Overview

Provide an overview of the topic to be studied...

1.3 Objectives

Explicitly state the objectives set to be achieved with this thesis...

1.4 Thesis Outline

Briefly explain the contents of the different chapters...

Chapter 2

Probabilistic Modelling

2.1 Introduction

Probabilistic modelling can be described as the task of finding the probability distribution that best describes a given dataset $\mathcal{D}$, i.e. a distribution $p(\mathcal{D})$ such that $\mathcal{D} \sim p(\mathcal{D})$ is as plausible as possible. Not only that, we normally want the distribution to be representative of the real process that generated $\mathcal{D}$. That is to say we want $p(\mathcal{D})$ to generalize to data we haven't yet observed.

There are effectively infinite possible distributions that could have generated the data we observe, but won't generalize to unobserved data. Commonly, the first way in which we restrict the set of possible distributions is to assume that the distribution of interest has a parametric form, and hence is defined by a set of parameters θ [1] .

Intuitively, the size of θ is deeply connected with the expressiveness of the model. In practice, this translates to the observation that if we make the model expressive enough, it can fit the observed data arbitrarily well. Naïvely, this would be a desirable characteristic. However, it normally comes at the cost of sacrificing generalization capability (TODO: Explain why). One way to counter this is to give preference to more parsimonious models - models that use less parameters and are less complex. There are strategies to make this mathematically objective and prevent the modeller from both having too complex and too simple models. Some of those methods are he Bayesian Information Criterion, the Akaike Information Criterion and the Minimum Description Length.

2.2 Structure and Latent Variables

In some cases, one might want to leverage some available domain knowledge. This often translates into assuming that there is some latent structure in the data. This structure is encoded into latent variables and their influence over the observable variables.

In this scenario, we become interested in the distribution given by $p(x, z, \theta_x, \theta_z)$, where z is the latent

[1] For the type of models and problems dealt with in this work, I will assume θ is finite, but it's worth noting that there are models which have θ *grow* with the dataset size. These are called non-parametric models. They come with their own advantages and disadvantages, which are out of the scope of this work.

variable, θ_x is the parameter vector for the distribution over $\mathcal{X}$, and θ_z is the parameter vector for the distribution over $\mathcal{Z}$.

For structure and latent variables to be useful we normally make the additional assumption that we have the ability of factoring that distribution in ways that make it tractable. One common factorization is:

$$p(x, z, \theta) = p(x|z, \theta_x)p(z|\theta_z)p(\theta_x)p(\theta_z) \tag{2.1}$$

,

2.3 Approximate Inference

For simplicity, let us consider θ as part of the latent variables z. This means that the model is simply written as the joint distribution: $p(x, z)$. *Inference*[2] is the task of finding the most probable z after having observed x . Specifically, the goal is to find the posterior distribution of z, given x, i.e.: $p(z|x)$.

Recall Bayes' Law:

$$\begin{align} p(z|x) &= \frac{p(x|z)p(z)}{p(x)} \tag{2.2} \\ &= \frac{p(x|z)p(z)}{\int p(x|z')p(z')dz'} \tag{2.3} \end{align}$$

For the vast majory of cases, the integral on the denominator will be intractable. To overcome this difficulty we normal resort to two families of methods: Monte-Carlo methods, and Variational methods.

2.3.1 Monte-Carlo Methods

Monte-Carlo methods work by using sampling techniques to approximate the intractable integral. The most powerful subclass of these methods is called Markov-chain Monte-Carlo. Its approach consists of devising a scheme that allows for sampling from a distribution close to the one of interest. It accomplishes this by defining a Markov-Chain whose transition function is guaranteed to make it converge asymptotically to the distribution of interest, given some constraints (ergodicity...) (TODO: explain more)

2.3.2 Variational Methods

Variational methods work by turning the problem of integration into one of optimization. They propose a family of parametric distributions, and then optimize the parameters so as to minimize the "distance" between the approximate (normally called "variational") distribution and the distribution of interest.

There are two ways to derive the most commonly used objective function for this problem.

[2] If we hadn't collapsed θ into z and were instead handling separately, we would call **inference** to the task of finding z and **learning** to the task of finding θ

Kullback-Leibler Divergence

The Kullback-Leibler divergence is a measure[3] of the distance between two probability distributions p, and q. It is given by:

$$KL(q||p) = \int q \log \frac{q}{p} \tag{2.4}$$

In the setting of inference, p is the posterior $p(z|x)$ and q is a distribution in some parametric family, with parameters ϕ, i.e., $q(z;\phi)$. However, it's clear that we can't compute the Kullback-Leibler directly, because it requires the knowledge of both the distributions, and finding $p(z|x)$ is precisely the task at hand. Let us expand the KL divergence expression:

$$\begin{aligned}
KL(q||p) &= \int q(z)(\log q(x) - \log p(z|x))dz \qquad (2.5)\\
&= \int q(z)(\log q(z) - (\log p(x,z) - \log p(x)))dz \qquad (2.6)\\
&= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(x,z)] + \mathbb{E}_q[\log p(x)] \qquad (2.7)\\
&= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(x,z)] + \log p(x) \qquad (2.8)
\end{aligned}$$

The last term is constant w.r.t $q(z)$. In that sense, for a fixed $p(x)$, minimizing the KL divergence is equivalent to minimizing $\mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(x,z)]$, which is equivalent to maximizing $\mathbb{E}_q[\log p(x,z)] - \mathbb{E}_q[\log q(z)]$. This quantity is commonly refered to as ELBO - Expectation Lower BOund. It can be rewritten as:

$$\begin{aligned}
ELBO(q) &= \mathbb{E}_q[\log p(x,z)] - \mathbb{E}_q[\log q(z)] \qquad (2.9)\\
&= \mathbb{E}_q[\log p(x|z)] + \mathbb{E}_q[\log p(x|z)] - \mathbb{E}_q[\log q(z)] \qquad (2.10)
\end{aligned}$$

In this form, each term of the ELBO has an easily interpretable role:

- $\mathbb{E}_q[\log p(x|z)]$ tries to maximize the conditional likelihood of x. That can be seen as assigning high probability mass to values of z that *explain* x well.

- $\mathbb{E}_q[\log p(x|z)]$ is the symmetric of the crossentropy between $q(z)$ and $p(x|z)$. Maximizing this quantity is equivalent of maximizing that crossentropy. This can be regarded as a regularizer that discourages $q(z)$ of being too different from the prior $p(z)$.

- $-\mathbb{E}_q[\log q(z)]$ is the entropy of $q(z)$. Maximizing this term incentivizes the probability mass of $q(z)$ to be spread out: another form of regularization.

[3] Note that the KL divergence isn't symmetric and as such I haven't called it a *metric*

A lower bound on $\log p(x)$

Another way of approaching the intractable posterior is to start by stating that our inherent goal is to maximize $p(x)$, or equivalently $\log p(x)$. Given that, consider the following:

$$\log p(x) = \log \int p(x,z) dz \Leftrightarrow \tag{2.11}$$

$$\Leftrightarrow \log p(x) = \log \int q(z) \frac{p(x,z)}{q(z)} dz \Leftrightarrow \tag{2.12}$$

$$\Leftrightarrow \log p(x) = \log \mathbb{E}_q[\frac{p(x,z)}{q(z)}] \Leftrightarrow \tag{2.13}$$

$$\Leftrightarrow \log p(x) \geq \mathbb{E}_q[\log \frac{p(x,z)}{q(z)}] \Leftrightarrow \tag{2.14}$$

$$\Leftrightarrow \log p(x) \geq \mathbb{E}_q[\log p(x,z)] - \mathbb{E}_q[q(z)] \Leftrightarrow \tag{2.15}$$

Chapter 3

Normalizing Flows

3.1 Introduction

Lorem ipsem Lorem ipsemLorem ipsemLorem ipsemLorem ipsem Lorem ipsemLorem ipsemLorem ipsemLorem ipsemLorem ipsemLorem ipsem Lorem ipsemLorem ipsemLorem ipsemLorem ipsem Lorem ipsemLorem ipsemLorem ipsemLorem ipsemLorem ipsem Lorem ipsemLorem ipsemLorem ipsemLorem ipsemLorem ipsem

Chapter 4

Variational Mixture of Normalizing Flows

4.1 Introduction

Lorem ipsem Lorem ipsemLorem ipsemLorem ipsemLorem ipsem Lorem ipsemLorem ipsemLorem ipsemLorem ipsemLorem ipsemLorem ipsem Lorem ipsemLorem ipsemLorem ipsemLorem ipsem Lorem ipsemLorem ipsemLorem ipsemLorem ipsemLorem ipsem Lorem ipsemLorem ipsemLorem ipsemLorem ipsemLorem ipsem

Chapter 5

Conclusions

Insert your chapter material here...

5.1 Achievements

The major achievements of the present work...

5.2 Future Work

A few ideas for future work...

Bibliography