



Leon Tiefenböck

# **Learning Probabilistic Circuits through sliced score matching**

## **Bachelors's Thesis**

to achieve the university degree of

Bachelor of Science

Bachelor's degree programme: Computer Science

submitted to

**Graz University of Technology**

## **Supervisor**

Ass.-Prof. Dipl.-Ing. Dr. techn. Robert Peharz

Graz, October 2024



# Abstract

Probabilistic Circuits (PCs) have shown to be a promising approach to probabilistic modelling, due to their many efficient and exact inference opportunities while still being complex enough to model arbitrary distributions. However they still are difficult to train and lack in results compared to other state of the art approaches in probabilistic modelling, especially with large high dimensional datasets. (Researchers fear that during the gradient based Maximum Likelihood optimization the algorithm gets stuck in local minima.) In this thesis I try to experimentally see if by using other learning objectives, namely sliced score matching, which recently gained prominence in learning energy based models (EBMs), we can work around these issues and achieve better results with PCs. I do this by training PCs using different algorithms on 2D and also high dimensional image data and comparing results.

# Contents

<b>Abstract</b>	<b>iii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Background</b>	<b>2</b>
2.1 Probabilistic Modelling . . . . .	2
2.2 Probabilistic Circuits . . . . .	3
2.3 Score Matching . . . . .	3
2.4 Sliced Score Matching . . . . .	3
<b>3 Methods</b>	<b>5</b>
<b>4 Experimental Results</b>	<b>6</b>
4.1 Dataset . . . . .	6
<b>5 Discussion</b>	<b>7</b>
5.1 Interpretation of Experimental Results . . . . .	7
<b>6 Conclusions and Future Work</b>	<b>8</b>
<b>Bibliography</b>	<b>10</b>

# 1 Introduction

## 2 Background

### 2.1 Probabilistic Modelling

At its very core Machine Learning (ML) aims to create algorithms/programs that learn from data in order to reason about it, make future predictions or perform all kinds of different inference tasks. One possible way to achieve this, especially when there is an inherent uncertainty in the data, which is the case for most real-world scenarios, is Probabilistic Modelling. In Probabilistic Modelling we use probabilities to express this uncertainty and the rules of probability theory for inference.

To make this more clear let's say we have some data and assume that this data was drawn randomly from some unknown distribution. We would call this distribution  $P^*$ . If this distribution was known to us all inference tasks would boil down to applying the basic rules of probability theory. The sum rule to compute marginals, the product rule to compute conditionals and more complex rules, derived from these two, to perform harder inference tasks.

However since we don't know  $P^*$ , we have to do something different, which leads us to a basic definition of probabilistic modelling: Through machine learning methods we create a framework that approximates an original unknown distribution, from which we only have limited samples, as closely as possible in order to reason about it with the rules of probability theory. [1].

Before continuing with different approaches to probabilistic modelling, I want to introduce two key concepts that are often used to discuss these different methods.

Definition 1: Expressiveness - a probabilistic model is called expressive if the learned distribution approximates the original distribution to a high degree. For instance this can be measured with Kullback-Leibler (KL) divergence [2], which measures how similar two probability distributions are. Two identical distributions would then have a KL-divergence of 0. In the Probabilistic Modelling case a KL-divergence of 0 would mean that the model is most expressive.

Definition 2: Tractability - a probabilistic model is called tractable with respect to an inference task, if the model can complete this task exactly and efficiently. Exactly in

this case means without relying on approximation e.g. Monte-Carlo Simulation and efficiently means in linear time with respect to the model size.

In recent years expressive capability has increased considerably through models based on neural networks like Generative-Adversarial Networks (GANs) [3], Variational Autoencoder (VAE) [4] and many others. However what these models excel in for instance generating very realistic images or compelling text in language modelling, they lack in tractability for all except the most simplest inference tasks. [1]

On the other hand there are many mostly older, less complex models like Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs) and so on, that lack in expressiveness and only work well enough for very simple data but can compute most if not all inference tasks tractably, which directly leads to Probabilistic Circuits.

## 2.2 Probabilistic Circuits

Recent progress in probabilistic modelling through .. have pushed the expressive capability (how close the approximated distribution is to the true one) ever forward. However what expressive gains these models achieve through their complexity they loose in tractability, the ability to perform inference tasks exactly and efficiently. More concretely tractability in one inference task means that a model can perform the task exactly, without approximation, and in polynomial time with respect to the model size. Except for the simplest tasks, these models must rely on costly approximation.

On the other hand older approaches like .. which are less complex and therefore lack in expressiveness can perform all sort of inference tasks tractably.

Probabilistic Circuits are a framework for probabilistic models that try to be both expressive and tractable. In essence they try to achieve this by only introducing complexity in a controlled manner. To be valid a PC must adhere to certain structural constraints.

## 2.3 Score Matching

## 2.4 Sliced Score Matching

Why score matching instead of minimizing NLL? should i even write about unnormalized models? how much detail for sm and ssm?

As described before, we want to learn the PDF of the data. In the case of PCs we directly model a normalized (integrates to 1) density but most other approaches model a unnormalized density. This is because the normalization constant is very costly or impossible to compute in these scenarios.

In score matching [5]

Which leads to this score matching objective:

However this introduces another problem. With high dimensional data the trace of the hessian is very costly to compute.

Here sliced score matching [6] is introduced. Instead of doing the costly hessian calculation the multidimensional data is projected onto  $n$  random vectors (slices). The authors in the paper show that ..

Which leads to this final sliced score matching objective:



## 3 Methods

## **4 Experimental Results**

### **4.1 Dataset**

## **5 Discussion**

### **5.1 Interpretation of Experimental Results**

## **6 Conclusions and Future Work**

# Appendix

# Bibliography

- [1] YooJung Choi, Antonio Vergari, and Guy Van den Broeck. “Probabilistic Circuits: A Unifying Framework for Tractable Probabilistic Models.” In: (Oct. 2020). URL: <http://starai.cs.ucla.edu/papers/ProbCirc20.pdf> (cit. on pp. 2, 3).
- [2] S. Kullback and R. A. Leibler. “On Information and Sufficiency.” In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86. ISSN: 00034851. URL: <http://www.jstor.org/stable/2236703> (visited on 09/12/2024) (cit. on p. 2).
- [3] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML]. URL: <https://arxiv.org/abs/1406.2661> (cit. on p. 3).
- [4] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML]. URL: <https://arxiv.org/abs/1312.6114> (cit. on p. 3).
- [5] Aapo Hyvärinen and Peter Dayan. “Estimation of non-normalized statistical models by score matching.” In: *Journal of Machine Learning Research* 6.4 (2005) (cit. on p. 4).
- [6] Yang Song et al. “Sliced Score Matching: A Scalable Approach to Density and Score Estimation.” In: (2019). arXiv: 1905.07088 [cs.LG]. URL: <https://arxiv.org/abs/1905.07088> (cit. on p. 4).