**FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University**

MASTER THESIS

Bc. Jakub Arnold

Bayesian Optimization of Hyperparameters Using Gaussian Processes

Institute of Formal and Applied Linguistics

Supervisor of the master thesis:	RNDr. Milan Straka, Ph.D.
Study programme:	Computer Science
Study branch:	Artificial Intelligence

Prague 2019

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date .... signature of the author

Dedication.

Title: Bayesian Optimization of Hyperparameters Using Gaussian Processes

Author: Bc. Jakub Arnold

Institute: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Milan Straka, Ph.D., Institute of Formal and Applied Linguistics

Abstract: Abstract.

Keywords: gaussian process bayesian optimization global optimization neural network

Contents

Introduction

0.1 Bayesian Optimization

Consider the problem of optimizing an arbitrary Lipschitz continuous function $f : \boldsymbol{\mathcal{X}} \to \mathbb{R}$ where $\boldsymbol{\mathcal{X}} \subset \mathbb{R}^d, d \in \mathbb{N}$. We call f the *objective function* and treat it as a black box. That is, we make no assumption on its analytical form, or on our ability to compute its derivatives. The main assumption is on its continuity in order to approximate it with a regression model. Our goal is to find the global minimum $\mathbf{x}_{\text{opt}}$ over the set $\boldsymbol{\mathcal{X}}$, that is

$$\mathbf{x}_{\text{opt}} = \arg\min_{\mathbf{x} \in \boldsymbol{\mathcal{X}}} f(\mathbf{x}).$$

We also assume that the evaluation of f is expensive, as the goal of Bayesian Optimization is to find the optimum as quickly as possible. If the function can be evaluated cheaply, other global optimization approaches such as simulated annealing or evolution strategies could potentially yield better results (TODO ref).

Let $\boldsymbol{\mathcal{D}}_n = \{(\mathbf{x}_i, y_i), i \in 1 : n\}$ denote a set of n samples (evaluations) of the function f, that is $y_i = f(\mathbf{x}_i)$. Our goal is to pick the next $\mathbf{x}_{n+1}$ to maximize our chance of finding the optimum quickly.

Consider the set of all continuous functions $f \in \boldsymbol{\mathcal{F}}$ with a prior distribution $p(f)$. Conditioning on our samples gives us a posterior distribution over possible functions $p(f|\boldsymbol{\mathcal{D}})$.

Conclusion

Bibliography

List of Figures

List of Tables

List of Abbreviations

A. Attachments

A.1 First Attachment