UNIVERSIDAD AUTÓNOMA DE MADRID

# Advanced Kernel Methods for Multi-Task Learning

by

Carlos Ruiz Pastor

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Escuela Politécnica Superior

Computer Science Department

under the supervision of José R. Dorronsoro Ibero

August 2021

*What is the essence of life? To serve others and to do good.*

Aristotle.

# *Abstract*

.

# *Resumen*

.

# *Acknowledgements*

.

# Contents

# Abbreviations

| | |
|---|---|
| **ADF** | **A**ssumed **D**ensity **F**iltering |
| **AF** | **A**cquisition **F**unction |
| **BO** | **B**ayesian **O**ptimization |
| **DGP** | **D**eep **G**aussian **P**rocess |
| **EI** | **E**xpected **I**mprovement |
| **EP** | **E**xpectation **P**ropagation |
| **GP** | **G**aussian **P**rocess |
| **KL** | **K**ullback **L**iebler |
| **MCMC** | **M**arkov **C**hain **M**onte **C**arlo |
| **PPESMOC** | **P**arallel **P**redictive **E**ntropy **S**earch for **M**ultiobjective **O**ptimization with **C**onstraints |
| **PES** | **P**redictive **E**ntropy **S**earch |
| **PESMOC** | **P**redictive **E**ntropy **S**earch for **M**ultiobjective **O**ptimization with **C**onstraints |
| **RS** | **R**andom **S**earch |
| **UCB** | **U**pper **C**onfidence **B**ound |

*To my family*

Chapter 1

# Introduction

**We begin this manuscript...**

## 1.1 Introduction

## 1.2 Publications

This section presents, in chronological order, the work published during the doctoral period in which this thesis was written. We also include other research work related to this thesis, but not directly included on it. Finally, this document includes content that has not been published yet and is under revision.

### Related Work

### Work In Progress

## 1.3 Summary by Chapters

In this section...

**Chapter 3** provides an introduction to GPs and the expectation propagation algorithm. Both are necessary concepts for the BO methods that we will describe in the following chapters. This chapter reviews the fundamentals of GPs and why they are so interesting for BO. More concretely, we review the most popular kernels, the analysis of the posterior and predictive distribution and how to tune the hyper-parameters of GPs: whether by maximizing the marginal likelihood or by generating samples from the hyper-parameter posterior distribution. Other alternative probabilistic surrogate models are also described briefly. Some of the proposed approaches of this thesis are extensions of an acquisition function called predictive entropy search, that is based on the expectation propagation approximate inference technique. That is why we provide in this chapter an explanation of the expectation propagation algorithm.

**Chapter 5** introduces the basics of BO and information theory. BO works with probabilistic models such as GPs and with acquisition functions such as predictive entropy search, that uses information theory. Having studied GPs in Chapter 3, BO can be now understood and it is described in detail. This chapter will also

describe the most popular acquisition functions, how information theory can be applied in BO and why BO is useful for the hyper-parameter tuning of machine learning algorithms.

**Chapter ??** describes an information-theoretical mechanism that generalizes BO to simultaneously optimize multiple objectives under the presence of several constraints. This algorithm is called predictive entropy search for multi-objective BO with constraints (PESMOC) and it is an extension of the predictive entropy search acquisition function that is described in Chapter 5. The chapter compares the empirical performance of PESMOC with respect to a state-of-the-art approach to constrained multi-objective optimization based on the expected improvement acquisition function. It is also compared with a random search through a set of synthetic, benchmark and real experiments.

**Chapter ??** addresses the problem that faces BO when not only one but multiple input points can be evaluated in parallel that has been described in Section **??**. This chapter introduces an extension of PESMOC called parallel PESMOC (PPESMOC) that adapts to the parallel scenario. PPESMOC builds an acquisition function that assigns a value for each batch of points of the input space. The maximum of this acquisition function corresponds to the set of points that maximizes the expected reduction in the entropy of the Pareto set in each evaluation. Naive adaptations of PESMOC and the method based on expected improvement for the parallel scenario are used as a baseline to compare their performance with PPESMOC. Synthetic, benchmark and real experiments show how PPESMOC obtains an advantage in most of the considered scenarios. All the mentioned approaches are described in detail in this chapter.

**Chapter ??** addresses a transformation that enables standard GPs to deliver better results in problems that contain integer-valued and categorical variables. We can apply BO to problems where we need to optimize functions that contain integer-valued and categorical variables with more guarantees of obtaining a solution with low regret. A critical advantage of this transformation, with respect to other approaches, is that it is compatible with any acquisition function. This transformation makes the uncertainty given by the GPs in certain areas of the space flat. As a consequence, the acquisition function can also be flat in these zones. This phenomenom raises an issue with the optimization of the acquisition function, that must consider the flatness of these areas. We use a one exchange neighbourhood approach to optimize the resultant acquisition function. We test our approach in synthetic and real problems, where we add empirical evidence of the performance of our proposed transformation.

**Chapter ??** shows a real problem where BO has been applied with success. In this problem, BO has been used to obtain the optimal parameters of a hybrid Grouping Genetic Algorithm for attribute selection. This genetic algorithm is combined with an Extreme Learning Machine (GGA-ELM) approach for prediction of ocean wave features. Concretely, the significant wave height and the wave energy flux at a goal marine structure facility on the Western Coast of the USA is predicted. This chapter illustrates the experiments where it is shown that BO improves the performance of the GGA-ELM approach. Most importantly, it also outperforms a random search of the hyper-parameter space and the human expert criterion.

**Chapter ??** provides a summary of the work done in this thesis. We include the conclusions retrieved by the multiple research lines covered in the chapters. We also illustrate lines for future research.

## 1.4 Definitions and Notation

Chapter 2

# Foundations and Concepts

**This chapter presents...**

## 2.1 Introduction

## 2.2 Kernels

### 2.2.1 Motivation and Definition

### 2.2.2 Reproducing Kernel Hilbert Spaces

### 2.2.3 Examples and Properties

## 2.3 Risk Functions and Regularization

### 2.3.1 Empirical and Expected Risk

### 2.3.2 Regularized Risk Functional

### 2.3.3 Representer Theorem

## 2.4 Optimization

### 2.4.1 Convex Optimization

### 2.4.2 Unconstrained Problems

### 2.4.3 Constrained Problems

## 2.5 Statistical Learning

### 2.5.1 Uniform Convergence and Consistency

### 2.5.2 VC dimension and Structural Learning

## 2.6 Support Vector Machines

### 2.6.1 Linearly Separable Case

### 2.6.2 Non-Linearly Separable Case

### 2.6.3 Kernel Extension

### 2.6.4 SVM properties

### 2.6.5 Connection with Structural Learning

### 2.6.6 SVM Variants

## 2.7 Conclusions

In this chapter, we covered...

Chapter 3

# Multi-Task Learning

**This chapter presents...**

## 3.1 Introduction

## 3.2 Why does Multi-Task Learning work?

### 3.2.1 Inductive Bias Learning Problem

Tipically in Machine Learning the goal is to find the best hypothesis $h(x, \alpha_0)$ from a space of hypothesis $\mathcal{H} = \{h(x, \alpha), \alpha \in \Lambda\}$, where $\Lambda$ is any set of parameters. This best candidate can be selected according to different inductive principles, which define a method of approximating a global function $f(x)$ from a training set: $z := \{(x_i, y_i),\ i = 1, \ldots, n\}$ where $(x_i, y_i)$ are sampled from a distribution $F$. In the classical statistics we find the Maximum Likelihood approach, where the goal is to estimate the density $f(x) = p(y \mid x)$ and the hypothesis space is parametric, i.e. $\mathcal{H} = \{h(x, \alpha), \alpha \in \Lambda \subset \mathbb{R}^m\}$. The learner select the parameter $\alpha$ that maximizes the probability of the data given the hypothesis. Another more direct inductive principle is Empirical Risk Minimization (ERM), which is the most common one. In ERM the densities are ignored and an empirical error $\hat{R}_z$ is minimized with the hope of minimizing the true expected error $R_F$, which would result in a good generalization. Several models use the ERM principle to generalize from data such as Neural Networks or Support Vector Machines. These methods are designed to find a good hypothesis $h(x, \alpha)$ from a given space $\mathcal{H}$. The definition of such space $\mathcal{H}$ define the bias for these problems. If $\mathcal{H}$ does not contain any good hypothesis, the learner will not be able to learn.

The best hypothesis space we can provide is the one containing only the optimal hypothesis, but this is the original problem that we want to solve. Therefore, in the single task scenario, there is no difference between bias learning and ordinary learning. Instead, we focus on the situation where we want to solve multiple related tasks. In that case, we can obtain a good space $\mathcal{H}$ that contains good solutions for the different tasks. In Baxter (2000) an effort is made to define the concepts needed to construct the theory about inductive bias learning, which can be seen as a generalization of strict multi-task learning. This is done by defining an environment of tasks and extending the work of Vapnik (2013), which defines the capacity of space of hypothesis, Baxter defines the capacity of a family of spaces of hypothesis.

Before presenting the concepts defined for Bias Learning, and to establish an analogy to those of ordinary learning, we briefly review some statistical learning concepts.

### Ordinary Learning

In the ordinary statistical learning, some theoretical concepts are used:

- an *input space* $\mathcal{X}$ and an *output space* $\mathcal{Y}$,
- a *probability distribution* $F$, which is unknown, defined over $\mathcal{X} \times \mathcal{Y}$,
- a *loss function* $\ell(\cdot, \cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and
- a *hypothesis space* $\mathcal{H} = \{h(x, \alpha), \alpha \in \Lambda \subset \mathbb{R}^m\}$ with hypothesis $h(\cdot, \alpha) : \mathcal{X} \to \mathcal{Y}$.

The goal for the learner is to select a hypothesis $h(x, \alpha) \in \mathcal{H}$, or equivalently $\alpha \in \Lambda$, that minimizes the expected risk

$$R_F(\alpha) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(h(x, \alpha), y)\, dF(x, y).$$

The distribution $F$ is unknown, but we have a training set $z = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of samples drawn from $F$. The approach is then is to apply the ERM inductive principle, that is to minimize the empirical risk

$$\hat{R}_z(\alpha) = \frac{1}{n} \sum_{i=1}^{n} l(h(x_i), y_i).$$

Although $\hat{R}_z(\alpha)$ is an unbiased estimator of $R_F(\alpha)$, it has been shown Vapnik (2013) that this approach, despite being the most evident one, is not the best principle that can be followed. This has relation with two facts: the first one is that the unbiased property is an asymptotical one, the second one has to do with overfitting. Vapnik answers to the question of what can be said about $R_F$ when $\alpha$ minimizes $\hat{R}_z(\alpha)$, and moreover, his results are valid also for small number of training samples $n$. More specifically, Vapnik sets the sufficient and necessary conditions for the consistency of an inductive learning process, i.e. for $\hat{R}_z(\alpha) \xrightarrow{P} R_F(\alpha)$ uniformly. Vapnik also defines the capacity of a hypothesis space and use it to derive bounds on the rate of this convergence for any $\alpha \in \Lambda$ and, more importantly, bounds on the difference $\inf_{\alpha \in \Lambda} \hat{R}_z(\alpha) - \inf_{\alpha \in \Lambda} R_F(\alpha)$. Under some general conditions, he proves that:

$$\inf_{\alpha \in \Lambda} \hat{R}_z(\alpha) - \inf_{\alpha \in \Lambda} R_F(\alpha) \leq C(n/d) \tag{3.1}$$

where $C$ is some non-decreasing function and $d$ is the capacity of the space $\mathcal{H}$, also named the VC-dimension $\mathcal{H}$. This means that the generalization ability of a learning process can be controlled in terms of two factors:

- The number of training samples $n$. A greater number of training samples assures a better generalization of the learning process.This looks intuitive and could be already inferred from the asymptotical properties.
- The VC-dimension $d$, or capacity, of the hypothesis space $\mathcal{H}$, which is desirable to be small. This term is not intuitive and is the most important term in Vapnik theory.

The VC-dimension measures the capacity of a set of hypothesis $\mathcal{H}$. If the capacity of the set $\mathcal{H}$ is too large, we may find a hypothesis $h\left(x, \alpha^{*}\right)$ that minimizes $\hat{R}_z$ but does not generalize well and therefore, does not minimize $R_F$. This is the overfitting problem. On the other side, if we use a simple $\mathcal{H}$, with low capacity, we could be in a situation where there is not a good hypothesis $h\left(x, \alpha\right) \in \mathcal{H}$, so the empirical risk $\inf_{\alpha \in \Lambda} R_F$ is too large. This is the underfitting problem.

### Bias Statistical Learning

In Baxter (2000) two main concepts are presented: the *family of hypothesis spaces* and an *environment* of related tasks. For a simpler formulation we consider $h(x) = h(x, \alpha)$ so we can substitute $\alpha$ by $h$ when necessary. Using these concepts, the bias learning problem has the following components:

- an *input space* $\mathcal{X}$ and an *output space* $\mathcal{Y}$,
- an *environment* $(\mathcal{P}, Q)$ where $\mathcal{P}$ is a set of distributions $P$ defined over $\mathcal{X} \times \mathcal{Y}$, and we can sample from $\mathcal{P}$ according to a distribution $Q$,
- a *loss function* $\ell\left(\cdot, \cdot\right) : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, and
- a *family of hypothesis spaces* $\mathbb{H} = \{\mathcal{H}_\delta, \delta \in \Delta\}$, where each element $\mathcal{H}_\delta$ is a set of hypothesis.

Analogous to ordinary learning, the goal is to minimize the expected risk, defined as

$$R_Q(\delta) = \int_{\mathcal{P}} \inf_{h \in \mathcal{H}_\delta} e_P(h) dQ(P) = \int_{\mathcal{P}} \inf_{h \in \mathcal{H}_\delta} \int_{\mathcal{X} \times Y} l(h(x), y) dP(x, y) dQ(P). \tag{3.2}$$

Again, we do not know $\mathcal{P}$ nor $Q$, but we have a training set samples from the environment $(\mathcal{P}, Q)$ obtained in the following way:

1. Sample $T$ times from $Q$ obtaining $P_1, \ldots, P_T \in \mathcal{P}$
2. For $r = 1, \ldots, T$ sample $m$ pairs $z_r = \{(x_1^r, y_1^r), \ldots, (x_m^r, y_m^r)\}$ according to $P_r$ where $(x_i^r, y_i^r) \in X \times Y$ .

We obtain a sample $z = \{(x_i^r, y_i^r), r = 1,\ i = 1, \ldots,\ m = 1, \ldots, T\}$, with $m$ examples from $T$ different learning tasks, and

$$\boldsymbol{z} := \begin{pmatrix} (x_1^1, y_1^1) & \ldots & (x_m^1, y_m^1) \\ \vdots & \ddots & \vdots \\ (x_1^T, y_1^T) & \ldots & (x_m^T, y_m^T) \end{pmatrix}$$

is defined as a $(T, m)$-sample. Using $\boldsymbol{z}$ we can define the empirical loss as

$$\hat{R}_{\boldsymbol{z}}(\delta) = \sum_{r=1}^{T} \inf_{h \in \mathcal{H}_\delta} \hat{R}_{z_r}(h) = \sum_{r=1}^{T} \inf_{h \in \mathcal{H}_\delta} \sum_{i=1}^{m} l(h(x_i^r), y_i^r), \tag{3.3}$$

which is an average of the empirical losses of each task. Note, however, that in this case this estimate is biased, since $R_{P_r}(h)$ does not coincide with $\hat{R}_{z_r}(h)$. To follow an analogous path to that of ordinary learning, the milestones in bias learning theory should include:

- Checking the consistency of the Bias Learning methods, i.e. proving that $\hat{R}_{\boldsymbol{z}}(\delta)$ converges uniformly in probability to $R_Q(\delta)$.
- Defining a notion of capacity of hypothesis space families $\mathbb{H}$.
- Finding a bound of $\hat{R}_{\boldsymbol{z}}(\delta) - R_Q(\delta)$ for any $\delta$ using the capacity of the hypothesis space family. If possible, finding also a bound for $\inf_{\delta\in\Delta} \hat{R}_{\boldsymbol{z}}(\delta) - \inf_{\delta\in\Delta} R_Q(\delta)$.

To address these points some previous definitions are needed. Two concepts that capture some important properties of $\mathbb{H}$ are introduced:

- The set of *sample-driven capacity* $\mathbb{H}^T$.
- The set of *distribution-driven capacity* $\mathbb{H}^*$.

In first place, he defines some concepts that allow to write two concepts of capacity related to $\mathbb{H}$. Using these two concepts, he shows uniform convergence of $e_Q(\mathcal{H})$ to $\hat{e}_z(h)$. Moreover, this result is valid not only asymptotically:

$$e_Q(\mathcal{H}) \leq \hat{e}_z(\mathcal{H}) + \epsilon$$

with $m$ such that

$$m \geq \frac{1}{T\epsilon^2} C(\mathbb{H})$$

where $C(\mathbb{H})$ is a notion of capacity of $\mathbb{H}$ (this bound is very simplified in this work). We can observe that as the number of tasks grow, we need less examples in each task. Also, if we have related tasks (and then $\mathbb{H}$ can have low capacity) the number of necessary examples per task also decreases. However, this bound is not so useful because it is theoretical, that is, we have theoretical notions of the capacity of $\mathbb{H}$ but we cannot compute it.

He then focuses on feature learning, that is, we define a specific type of learners where we have two steps:

- Learning the space of features (or learning $\mathcal{H}_f$)
- Learning the best linear model $g$ over the features given by $f$ (or learning the best hypothesis $g \circ f$ in $\mathcal{H}_f$)

For these type of models, he also shows how a multi-output Neural Network (which can be seen an MTL NN) is a model that embodies this feature learning approach. The NN can be seen as a linear model $g$ over a transformation $f$ of the original data. That is, the set of functions $\mathcal{H}_f = \{g \circ f\}$ used by the neural network is learned trying to minimize the empirical loss when we update $f$. Baxter derives specific bound for the NN using that the capacity of the hypothesis space family $\mathbb{H}$ can be computed for this case.

Finally, he show some results for Multi-Task Learning. Multi-Task Learning is a particular case of Bias Learning where we are not interested in finding an optimal hypothesis space $\mathcal{H}$ where we can find good solutions for a great variety of problems, but our goal is to find good solutions for a fixed number of tasks. That is, instead of having a family of probabilities $\mathcal{P}$ with a probability measure $\mathcal{Q}$, we have a fixed vector $\boldsymbol{P} = (P_1, \ldots, P_T)$ of probability measures over a fixed set of tasks and the learner minimizes:

$$e_{\boldsymbol{P}}(\mathcal{H}) = \sum_{r=1}^{T} e_{P_r}(h_r) = \sum_{r=1}^{T} \int_{X\times Y} l(h_r(x), y) dP_r(x, y). \tag{3.4}$$

Then, we define the empirical multi-task error as

$$\hat{e}_z(\mathcal{H}) = \sum_{r=1}^{T} \hat{e}_{z_r}(h_r) = \sum_{r=1}^{T} \sum_{i=1}^{m} l(h_r(x_i^r), y_i^r). \tag{3.5}$$

The difference is subtle but we have a different objective. In the multi-task setting, the hypothesis space $\mathcal{H}$ found will not necessarily be good for generalizing to new tasks. Also, we have got rid of the inf term, which was difficult to deal with.

Restricting to the Boolean functions, Baxter defines an extension of the VC-dimension $d_{\mathbb{H}}(T)$ that can be computed (for these Boolean map functions). Using that, he extends the previous theorems obtaining "computable" bounds and he derives a lower bound for $m$, if $m$ is below that lower bound then we can find tasks such that the error made by our estimator is larger then certain $\epsilon$.

### 3.2.2 A notion of Task relatedness

### 3.2.3 Learning Under Privileged Information

## 3.3 Multi-Task Learning Methods: An Overview

### 3.3.1 Bias Learning Approach

### 3.3.2 Low-Rank Approach

### 3.3.3 Learning Task Relations

### 3.3.4 Decomposition Approach

## 3.4 Deep Multi-Task Learning

### 3.4.1 Hard Parameter Sharing

### 3.4.2 Soft Parameter Sharing

## 3.5 Multi-Task Learning with Kernel Methods

## 3.6 Conclusions

In this chapter, we covered...

Chapter 4

# A Convex Formulation for Regularized Multi-Task Learning

## 4.1 Introduction

## 4.2 Convex Multi-Task Learning Support Vector Machines

### 4.2.1 Convex Formulation

### 4.2.2 L1 Support Vector Machine

### 4.2.3 L2 Support Vector Machine

### 4.2.4 LS Support Vector Machine

## 4.3 Optimal Convex Combination of trained models

## 4.4 Experiments

## 4.5 Conclusions

In this chapter, we have...

Chapter 5

# Adaptive Graph Laplacian Multi-Task Support Vector Machine

## 5.1 Introduction

## 5.2 Graph Laplacian Multi-Task Support Vector Machine

## 5.3 Adaptive Graph Laplacian Algorithm

## 5.4 Experiments

## 5.5 Conclusions

In this chapter, we have...

# Bibliography

Baxter, J. (2000). A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.