Dipartimento di Informatica e
Scienze dell'Informazione

Regularization Approaches in Learning Theory

by

Lorenzo Rosasco

Theses Series **DISI-TH-2006-05**

DISI, Università di Genova
v. Dodecaneso 35, 16146 Genova, Italy http://www.disi.unige.it/

Università degli Studi di Genova

Dipartimento di Informatica e
Scienze dell'Informazione

Dottorato di Ricerca in Informatica

Ph.D. Thesis in Computer Science

Regularization Approaches in Learning Theory

by

Lorenzo Rosasco

April, 2006

Dottorato di Ricerca in Informatica
Dipartimento di Informatica e Scienze dell'Informazione
Università degli Studi di Genova

DISI, Univ. di Genova
via Dodecaneso 35
I-16146 Genova, Italy
http://www.disi.unige.it/

Ph.D. Thesis in Computer Science (S.S.D. INF/01)

Submitted by Lorenzo Rosasco
DISI, Univ. di Genova
rosasco@disi.unige.it

Date of submission: April, 3 2006

Title: Regularization Approaches in Learning Theory

Advisors: Alessandro Verri
DISI, Univ. di Genova
verri@disi.unige.it

Ernesto De Vito
Dipartimento di Matematica, Univ. di Modena e Reggio Emilia
Via Campi 213/B, I-41100 Modena, Italia
devito@unimo.it

Supervisor: Alessandro Verri
DISI, Univ. di Genova
verri@disi.unige.it

Ext. Reviewers: Tomaso Poggio
CSAIL, Brain Sciences Dept., MIT, 46-5177B
43 Vassar Street, Cambridge, MA 02142, USA
tp@ai.mit.edu

Partha Niyogi
Dept. of Computer Science, Univ. of Chicago
1100 E. 58th Street, Ryerson Hall, Room 167, Hyde Park,
Chicago, IL 60637, USA
niyogi@cs.uchicago.edu

Abstract

Learning from examples can be seen as a very general framework for modeling a variety of different statistical inference problems. Such statistical problems are at the basis of the design of programs which are trained, instead of programmed, to perform a task. In particular supervised learning aims at finding an unknown input-output relation given a (possibly small) number of input-output instances (the examples). The main goal in this setting is not to describe the available data but to predict the output when a new input is given, that is to be able to generalize. A learning algorithm should be able to avoid over-fitting the data that is to over-estimate the importance of the available information loosing generalization properties. Regularization Theory was originally developed and formalized as a way to find stable solutions to ill-posed problems. Eventually some regularization techniques became popular in the context of machine learning as an effective way to avoid over-fitting. Though the parallel between learning and regularization for inverse problems has been proposed many years ago, a more formal connection between the two theories has not been pursued. In our study we show that in fact learning with a quadratic cost function defines a suitable inverse problem. The byproduct of this fact is that we are able to adapt many tools from regularization of inverse problems to learning theory. In particular we propose new algorithms and study their theoretical properties by means of novel proof techniques. As we consider more general cost functions the connection with inverse problems becomes less immediate but nonetheless we can use many of the developed theoretical tools to give a compact and unifying description of a large class of algorithms, namely regularization networks.

To my mother and to my father

"Non vi é progresso, non vi é rivoluzione di evi, nella vicenda del sapere, ma al massimo continua e sublime ricapitolazione..." (venerabile Jorge de Burgos, ne "Il nome della rosa", Umberto Eco)

Acknowledgements

I want to thank .....

Table of Contents

Chapter 1

Introduction

In this thesis we study an approach to the problem of learning which is based on the connection with the theory of regularization for ill-posed inverse problems. We start by explaining and motivating our point of view.

1.1 Learning is an Ill-Posed Problem

First we try to give an idea of what is learning from examples. A definition which is useful to explain our point of view is given in [Val84] "We say that a program for performing a task has been acquired by learning if it has been acquired by any means other then explicit programming." In other words, learning from examples refers to systems that are *trained*, instead of *programmed*, to perform a task. The training is based on a set of *examples* of the task we want to learn. Let us describe a few problems that can be addressed and has been addressed within the learning from examples paradigm.

Example 1 (Image Classification). *The problem of content-based image classification loosely refers to associating a given image to one of a predefined set of classes using the visual information contained in the image. There are several well known computer vision problems which can be viewed as instances of image classification. In particular we recall the problem of object detection, where the goal is finding one or more instances of an object in an image; or the problem of visual categorization which refers to associating an image to two or more image categories (e.g. indoor/outdoor or other higher level image descriptions).*

Example 2 (Text Categorization). *Another example of content based information retrieval is text categorization. In this case the goal is labeling texts document according to*

a predefined set of semantic categories. The importance of such a task can be easily understood thinking of the flow of documents in the World Wide Web which requires to be indexed and organized.

Example 3 (Classification of Microarray Data). *Finally an interesting application can be seen in the the context of the analysis of genomic data. In particular a problem that gained considerable attention is that of classifying genes using gene expression data from DNA micro-array hybridization experiments. Micro-array technology has provided biologists with the ability to measure the expression levels of thousands of genes in a single experiment. The vast amount of raw gene expression data leads to statistical problems such as the classification of the dataset into a collection of predefined classes. The goal in this case is often, not only to have few classification errors, but also to identify the differentially expressed genes that may be used to predict class membership for new samples. A main aspect in of micro-array classification is that usually a very small number of samples is available, compared to the number of genes in the sample, moreover there is often a big experimental variation in measured gene expression levels.*

In the above examples we focused on a specific type of learning problem, the so called supervised learning, which we will study in our analysis. A common aspect in the above problems is that the data can be thought of as input-output pairs, the examples. *The goal is, given a new input, to predict its output.* Clearly such a task is hopeless if we do not assume some model relating input and output, that is a model for the data. Once we agree on a model, learning proceeds similarly to natural science, that is through an inference process.
Then a supervised learning problem can be described as: given a certain (possibly small) number of observations we want to recover an approximation of the model underlying them. The problem is not trivial since we always have a finite amount of information available and various causes of uncertainty might affect the problem. In other words the problem is ill-posed [Had02, Had23]: in particular the solution is not unique and can change dramatically if we slightly change the data. In natural science one often has strong prior information which induces constraints on the possible solutions, while a main feature of learning is trying to minimize the amount of prior assumptions. Intuitively, we can say there is a trade-off between prior information and the amount of data available.
Given the above premises the question is then how we can design algorithms for learning and what are the main factors determining their performances.
A (simplified) graphical visualization of a 2-dimensional toy problem is useful. In fig. 1.1 each input point is a 2-dimensional vector and its label is given by its shape (triangle or circle). On the top left we see a set of data which can be thought of as a sample from a larger (possibly infinite) population on the top right. In this model the goal is then two draw a line such that points belonging to different class falls in different sides. It is crucial to remember the goal is not only to describe the available data but rather to be

predictive on new data. A solution which perfectly separates the data (bottom left) can perform poorly on other point of the same population (bottom right). Even in this toy

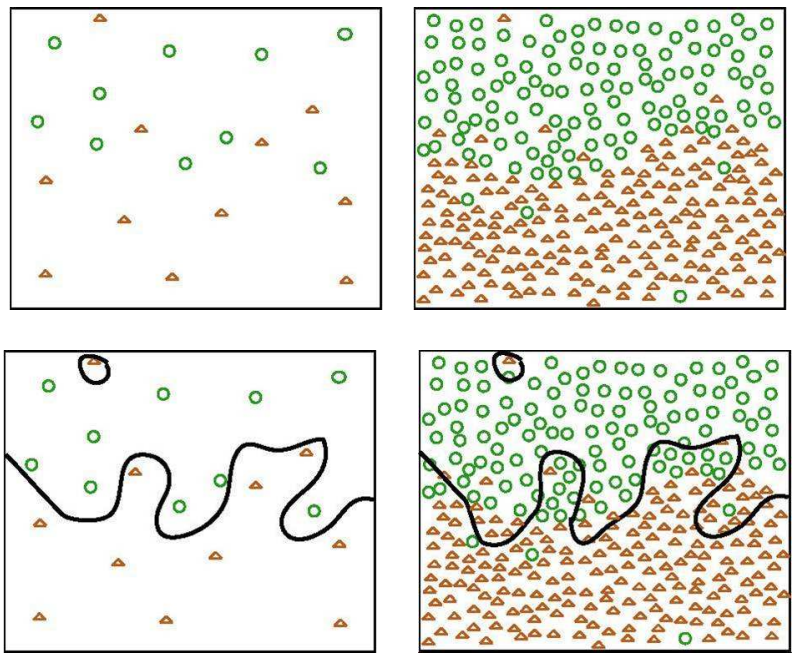

Figure 1.1: A 2-dimensional toy classification problem. Top left there is a data sample from a larger population on top right. Bottom left we can see a solution with zero empirical error and bottom right we see the poor performance of the same solution on the entire population.

model we can see some features of the problem. If we simply try to find a prediction rule which performs well on the data we tend to perform bad on new data. Clearly this is due to the fact that we have a finite number of examples. Moreover it should be clear that the more complex is the problem at hand the more examples we need: we can see an interplay between regularity of the target and number of required data.

1.2 Algorithms for Learning

From the examples of the previous section we learned that if the problem is difficult enough, a solution which simply describes the data is too irregular or too complex, that is it is an over-fitting solution. If we postulate that the problem has some regularity properties, then we might want to impose constraints on the class of possible solutions. As we will discuss in the following, this is exactly one of the fundamental results in learning theory. Regularity

is described in terms of complexity (see [Bou02] and references therein) or stability of the solution (see [PRMN04] and references therein).
The idea is then to write algorithms taking into account the following two quantities:

- an empirical error on the data that we indicate with ERR;
- a penalization term, PEN, that adds regularity constraints on the solution we are looking for. Clearly, the way such a term is designed should depend on the regularity assumptions on the problem.

If we develop this line of reasoning two questions arise naturally. The first one is:
"How can we design algorithms taking into account (explicitly or implicitly) both ERR and PEN?"
A possible way of doing this is simply considering objective functions of the form

$$ERR + \lambda PEN \tag{1.1}$$

where we introduce a positive parameter λ, the regularization parameter, which allows us to trade-off the fitting term and the complexity/smoothness term. We note that the introduction of the regularization parameter λ is crucial: taking $\lambda = 0$ we just take into account the error on the data, whereas taking λ big enough we forget about the empirical error to choose a very regular solution. The regularization parameter λ enables us to pass from over-fitting to over-smoothing and choosing it correctly allows us to prevent both problems.
This leads to the second question:
"How do we decide the correct trade-off between fitting the data and finding a regular solution?"
or in other words
"How do we choose the regularization parameter λ?"
Before discussing our approach to the above questions we comment on them. The point of view of balancing out complexity and empirical error is widely recognized as a crucial ingredient to develop learning algorithms. The expression (1.1) describes many learning algorithms, solving a Lagrangian minimization problem, such as Regularization Networks [EPP00]. Support Vector Machines [Vap98] are probably the most prominent example in this class of algorithms. Anyway there exists algorithms implementing the trade-off between ERR and PEN in other ways than (1.1). For example greedy algorithms inspired to boosting [MR03]. In any case for each algorithm there is at least one parameter to choose controlling its performance. Actually in practice more than one parameter is considered and this is more difficult both from a practical and a theoretical point of view. Choosing the parameters values requires to understand, not only of the meaning of each parameter, but also of the interplay between them. For example some parameters might depend on each other or have an influence on the solution which is negligible with respect to some

others. For any given algorithm we should first consider how many parameters we have to choose and what is their meaning. Even more important is the question of how to choose the regularization parameter. This is arguably one of the most important issues in learning since the regularization parameter choice completely determines the performance of a given algorithm. If an algorithm is very stable with respect to the parameter choice, this usually indicates that the problem at hand is simple and that even if we consider very different solutions we can get fairly good results. In general this is not the case and we expect the theory to provide methods for the regularization parameter choice. Unfortunately there is a huge gap between theory and practice. The theory so far can (at most) *explain* the meaning of the regularization parameters and give *qualitative* indications on the way to choose them. In practice, heuristics are used to choose the parameters with no theoretical guarantees. We note that heuristics are tricky in learning because we are dealing with an inference problem and we have a limited amount of data. This makes very difficult to design experimental protocols ensuring that the performance of a certain heuristics is reliable and reproducible. In other words, many theoretical results, important to answer extremely practical questions, are still missing.

The above discussion motivates a more abstract approach towards a better understanding and formalization of the properties of the algorithms.

For the sake of simplicity we restrict ourselves to a setting such that:

- each algorithm depends on one regularization parameter, so that in fact it provides us with a one parameter family of possible solutions;
- the final solution is then obtained defining a suitable regularization parameter choice.

A complete algorithm is then a one parameter family of algorithms provided with some regularization parameter choice. This is essentially the point of view of regularization theory for ill-posed inverse problems. In the following section we develop this last observation which is one of the main theme of this thesis. Interestingly the scheme we described above is also very similar to those considered in mathematical statistics as we discuss in the following remark.

Remark 1 (Regularization Model Selection). *The kind of framework we described is also related to the theory of model selection (see for example [BBM99]). In this context one can minimize the error on a discrete collection of solution spaces (the models) and has to pick up the model giving the best solution. First, we have to define a class of models and then we have to select the best model. In our approach, rather than a structure of solution spaces, we think in terms of family of algorithms (or maps). The problem of selecting the best model is then intimately related to the problem of choosing the regularization parameter.*

1.3 Is Learning an Inverse Problem?

Inverse problem theory deals with the following problem: observing the effects of a certain process, we want to recover the causes that generated them. According to the above informal definition, learning can indeed be seen as an inverse problem: given data generated by some model we would like to invert the process and recover information on the model itself. Anyway the theory of inverse problems is an established mathematical theory which has applications in a variety of different fields (see for example [BB98]). An inverse problem has a precise mathematical formulation in a functional analytical framework and typically involves the inversion of a linear equation. Such an inversion operation is often ill-posed. According to Hadamard [Had02, Had23] a problem is well posed if it has a unique solution, depending continuously on the data and it is ill-posed if it is not well-posed. The theory of Regularization provides well-posed approximation to ill-posed inverse problems. The intuition behind regularization is strikingly similar to the ideas we discussed in the previous sections. To avoid oscillatory behavior of the solution for small changes of the data, regularization constraints the regularity of the solution to achieve a smooth dependence to the data. It should not surprise that some regularization algorithms, such as the well known Tikhonov regularization [TA77], have exactly the form (1.1). Nonetheless in the context of inverse problems the idea of regularization has been mathematically formalized and the problem of regularization parameter choice is known to be central.

The above analogies motivate a better understanding of the relation between learning from examples and inverse problems. Hopefully a precise connection between the two theories would allow us to transfer results and tools from one theory to the other. In particular we might expect the theory of inverse problems to be helpful to:

- define new learning algorithms and shed lights to the properties of existing ones;
- define new principled way of choosing the regularization parameter;
- import theoretical results, concepts and proof techniques.

Despite the philosophical and algorithmic analogies, the task of finding a quantitative connection between learning theory and the theory of inverse problems is not straightforward. In fact, the setting where the two theories are developed are at first sight very different. One of the main outcome of our analysis is that learning can indeed be translated into an inverse problem. This allows us to draw a deep and useful connection and we review the main results that we derived in the following section.

1.4 Contributions

This section summarizes the main contributions of this thesis.

- **Learning and Inverse Problems** We give a rigorous study on the connection between learning and inverse problems highlighting similarities and differences between the two theories. If we consider quadratic cost and hypotheses space which are Reproducing Kernel Hilbert Spaces we can show that learning from a given sample of data defines a linear inverse problem. Such a problem can be seen as a stochastic discretization of an infinite dimensional inverse problem defined by a linear embedding equation. The latter is indeed the problem we *want* to solve, whereas the former is the problem we *can* solve. In classical inverse problems discretization and noise are treated separately since the discretization can be controlled[1]. This is not the case in learning the sampling of data has a probabilistic nature and requires the definition of new perturbation measures. Initially the above framework is applied in the study of Tikhonov regularization giving new consistency proofs. The analysis of Tikhonov regularization started in [DVCR05] is later refined in [DVRC$^+$05b, DVRC05a]. The connection between learning and inverse problem is considered in [DVRC$^+$05b, DVRCG04] and extended to a more general setting in [DVRC05a].

- **Regularization for Learning** A direct consequence of the analysis of learning and inverse problems is that we can consider many other algorithms besides Tikhonov regularization. The gradient descent algorithm, related to Landweber regularization, is studied in [YRC05], where early stopping regularization is shown to be consistent and the connection with boosting algorithm is discussed. In [RDVV05] and [BPR05] we considered a more general framework allowing to deal with a large class of algorithms in a unified way. On one hand we gave an abstract characterization of regularization for learning and prove various theoretical results such as finite sample bounds both for regression and classification. On the other hand we could show that instances in the proposed class of algorithms defines new kernel methods which have different features and are easy to implement. Finally we proposed a data driven choice for the regularization parameter achieving adaptivity in reproducing kernel Hilbert spaces.

- **Tikhonov Regularization with Convex Loss** When we consider cost function different from the quadratic one we loose the linear structure the made the comparison with inverse problems easier. For this reason we restrict our analysis to convex loss functions and to regularization schemes inspired by Tikhonov regularization. Indeed the convexity assumption is very natural both from the theoretical and the practical

[1] In fact discretization can have a stabilizing effect [EHN96].

view point since, on one hand it allows to use many tools form convex analysis and on the other hand it allows to define algorithms which are computationally efficient. For the considered class of algorithms we studied existence, uniqueness and, especially, the explicit form of the regularized solution [DVRC$^+$04, CR$^+$05]. In particular we consider the case when the penalization is a semi norm and derive a new quantitative version of the representer theorem [KW70] which express the solution in a closed form. This last result covers both the case of finite number of data (sample case) and the case of infinite number of data (population case) and allows to give a new prove of consistency for regularized kernel methods.

The main inspiration for our study are the works of [PG92, GJP95] (and eventually [NG99, EPP00, MNPR04]) and the work of [CS02b] (see also [CS02a, SZ04, SZ05]). The former works suggest that learning can be seen as a multivariate function approximation problem in which regularization tools can be fruitfully used. Such an observation was eventually helpful to highlight an intimate relation between regularization ideas and learning algorithms such as Support Vector Machines originally introduced as large margin algorithms. Though a complete mathematical connection between learning and inverse problem is not given the connection between the two theories is made explicit. On the other hand the work of [CS02b] gives a formulation of learning in functional analytical framework which is ideal to investigate the connection with the theory of inverse problems. The results in [CS02b] characterize the property of learning algorithms by means of complexity measures such as covering numbers. Developing on such results, in [DVCR05] we proposed an approach replacing the complexity analysis with random operators estimates. Such an approach is indeed at the basis of many of our later developments.

1.5 Structure of the Thesis

We end this chapter briefly describing the structure of the thesis.

In chapter 2 we first discuss the framework of supervised learning stating and commenting on the main mathematical assumptions. Then we recall an approach to the analysis of the algorithm based on the minimization of the empirical error on a ball in a reproducing kernel Hilbert space [Aro50].

In chapter 3 we discuss in detail the connection between learning and inverse problems. In particular:

- We show that learning is described by a suitable inverse problem and its stochastic discretization.

- We define new perturbation measures which are based on concentration inequalities for Hilbert space valued random variables.
- We give an abstract definition of regularization and describe in detail several kernel methods falling in this definition.
- Finally we discuss how we can describe the regularity of a problem using well known concept form the theory of inverse problems.

In chapter 4 we discuss the theoretical properties of the considered class of regularization methods. For the class of methods previously introduced we discuss finite sample bounds and consistency both for regression and classification.

In chapter 5 we consider Tikhonov regularization with a convex loss function. First we present the main result which is a new version of the representer theorem holding both for the sample case and population case. Then we study in some detail the meaning of regularization when the penalty term is a semi-norm. After discussing the issue of existence and uniqueness we show that many proposed algorithms are described within the framework we considered. Using the above results we recall a new approach to consistency of such algorithms and discuss their application in classification.

Finally in chapter 6 we give with a short summary of the presented analysis and discuss some open problems.

Chapter 2

Learning from Examples

In this section we introduce the main concepts, assumptions and notations describing the framework we consider as our model for learning from examples. We refer to [CS02b, Vap98, EPP00, SS02, BBL04a] and references therein for a broader introduction. The plan of the chapter is as follows. In section 2.1 we give a brief informal overview of the main concepts of supervised learning. In section 2.2 we discuss in detail the mathematical framework we consider. In 2.3 we give an overview of the theory of reproducing kernel Hilbert spaces. In section 2.4 we give our results on the analysis of empirical risk minimization (ERM) algorithm and in section 2.5 we discuss how ERM leads to regularization.

2.1 Supervised Learning at a Glance

We first give a description of the learning problem that we formalize in the following sections.

As we previously mentioned the idea of (supervised) learning is to infer an unknown input-output relation on the basis of a given set of input-output instances. The available data, namely the training set, are a collection of couples $\mathbf{z} = (x_1, y_1), \ldots, (x_n, y_n)$ where we can think of x as vectors and y as numbers. The aim of learning is to find a function $f_{\mathbf{z}}$ based on the data such that given a new input x_{new}, $f_{\mathbf{z}}(x_{new})$ is a *good* estimate of the output y_{new}. Before we can proceed any further defining the problem (i.e. clarifying what we mean by "good" estimate) we need to specify a model for the data.
To allow modeling the uncertainty in the learning process we assume that input and output are related by a probabilistic relation $\rho(x, y) = \rho_X(x)\rho(y|x)$. The training set $(x_1, y_1), \ldots, (x_n, y_n)$ is sampled according to ρ. Immediately we observe some facts. First rather than finding an estimator which is describing the available data we aim at finding

an estimator which is able to be descriptive of *new* data, namely which is able to *generalize*. Second there's not a single output associated to an input but a whole distribution of values: any possible solution will make some *errors*.
This motivates the definition of the *loss function* $\ell(y, f(x))$ quantifying the error we make predicting $f(x)$ when in fact the real output is y. Once we have chosen a loss functions the generalization property of a given function can be better formalized introducing the *expected error* (or *expected risk*)

$$\mathcal{E}(f) = \int_{X \times Y} \ell(y, f(x)) d\rho(x, y)$$

which can be seen as the average error of the function f on all the possible (x, y) couples weighted with respect to the probability measure ρ. *It should be clear that in order to have good generalization property an estimator should have small expected error* $\mathcal{E}(f_{\mathbf{z}})$.
Such a statement deserves some care in fact, since $f_{\mathbf{z}}$ (and hence $\mathcal{E}(f_{\mathbf{z}})$) depends on the data, the expected error of an estimator is probabilistic in nature and we need to clarify what kind of probabilistic analysis we are interested in. For example it is common in statistics to require the expected risk of the estimator to be small in expectation, that is we wish to control

$$\mathrm{E}[\mathcal{E}(f_{\mathbf{z}})].$$

In Learning Theory it is preferred a worst case analysis studying, for all $\varepsilon > 0$, the tail

$$\Pr\left(\mathcal{E}(f_{\mathbf{z}}) \geq \varepsilon\right),$$

of the expected risk.

Since we do not know the probability ρ but just the training set, we have no access to the expected error so that it is natural to ask how we can build an estimator achieving good generalization properties and what are the main factors affecting its performance. Indeed this will be the central theme in our studies.

2.2 The Ingredients

In the following sections we develop in details the mathematical and conceptual framework we use for our analysis.

2.2.1 Sample Space

We denote with X the *input space* and with Y the *output space*. Moreover we call *sample space* the Cartesian product $Z = X \times Y$. In the following we assume that X is a complete

separable metric space, that is a Polish space. Moreover we assume that the output space Y is a subset of $\mathbb{R}$ for regression and $\{-1, 1\}$ for classification. It follows that the sample space Z is also a Polish space. We model the input-output relation endowing Z with a probability measure ρ. We denote with ρ_X the marginal distribution on X, if $\pi : Z \to X$ is the projection from the sample space into the input space then $\rho_X = \rho \circ \pi^{-1}$. Moreover if $\mathcal{B}(Y)$ are the Borel subsets of $\mathbb{R}$, let $\rho(\cdot|\cdot) : \mathcal{B}(Y) \times X \to [0, 1]$ be a regular conditional probability, such that

$$\int_{X\times Y} g(x, y)d\rho(x, y) = \int_X d\rho_X(x) \int_Y d\rho(y|x)g(x, y)$$

for all measurable functions $g : X \times Y \to [0, \infty]$. Existence of both measures is ensured since Z is a Polish space [Dud02].
The probability measure ρ is fixed but unknown and the available data are a set of *examples*, that is a sample of n pairs (x, y) *identically and independently distributed (i.i.d.)* according to ρ. In the following let $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = (x_1, y_1), \ldots (x_n, y_n)$ be the *training set* which is the collection of available examples. Under the i.i.d. assumption we have that $\mathbf{z}$ belongs to the space Z^n endowed with the product measure $\rho^{\otimes n} = \rho \otimes \cdots \otimes \rho$. In the following we often denote such measure with P.
We call *estimator* a function $f_{\mathbf{z}} : X \to \mathbb{R}$ built on the basis on the given training set and *learning algorithm* any map $\mathbf{z} \to f_{\mathbf{z}}$.

We end this section adding a few comments on the assumptions we consider on the data space. First, it is common to take $X \subseteq \mathbb{R}^d$. In doing so we implicitly take into account a first representation of the input objects (due for example to the acquisition procedure) or a default representation procedure. Since the representation is undoubtedly one of the central concepts in learning we prefer to distinguish it from the notion of input space. This is somewhat related to fuzzy difference between variables and features [GP03]. To avoid this confusion we let the input space X be defined in an abstract way as a Polish space which is, mathematically, the natural notion of space to consider. Second, the input space X is often assumed to be compact, indeed this is just a technical assumption which is often not needed. Similarly it is often assumed $Y = [-M, M]$, for some $M \in \mathbb{R}^+$. This boundedness condition (clearly verified in classification) is a technical assumption, useful in some probabilistic analysis, which can be sometimes relaxed to some weaker assumption. Third, we note that most results we present in the following hold, with minor modifications, if we take Y to be a more general space, for example $\mathbb{R}^d$ or even a Hilbert space [DVC05]. Indeed this might be interesting as one wish to consider regression problem for vector valued functions [MP05b]. For sake of clarity in the following we restrict ourselves to the case $Y \subseteq \mathbb{R}$.

2.2.2 Loss function, Expected and Target Function

As we already discussed since we look for a deterministic rule in a probabilistic setting any possible solution will make some errors. This leads to the definition of *loss function* $\ell(y, f(x))$ which is a non-negative map $\ell : Y \times \mathbb{R} \to [0, +\infty[$ which is a point-wise error measure. We note that usually the loss functions depends on the difference $y - f(x)$ in regression and on the product $yf(x)$ in classification. A list of loss functions commonly in use include

- the square loss $\ell(w, y) = (w - y)^2$,
- the absolute value loss $\ell(w, y) = |w - y|$,
- the $\varepsilon-$insensitive loss $\ell(w, y) = \max\{|w - y| - \varepsilon, 0\} =: |w - y|_\varepsilon$

for regression, and

- the square loss $\ell(w, y) = (w - y)^2 = (1 - wy)^2$,
- the hinge loss $\ell(w, y) = \max\{1 - wy, 0\} =: |1 - wy|_+$,
- the logistic loss $\ell(w, y) = \ln(1 + e^{-wy})$

for classification. We often need some more restrictions on the form of the loss function. We collect the main mathematical assumptions in the following definition and then comment on the purpose of each assumption.

Definition 1. *Given $p \in [1, +\infty[$, a loss function is called* admissible *with respect to ρ if the following conditions hold true*

1. *for all $y \in Y$ the function $\ell(y, \cdot)$ is convex on $\mathbb{R}$;*
2. *the function ℓ is measurable on $Y \times \mathbb{R}$;*
3. *there are $b \in [0, +\infty[$ and $a : Y \to [0, +\infty[$ such that*

$$\ell(y, w) \leq a(y) + b|w|^p \qquad \forall w \in \mathbb{R},\ y \in Y \tag{2.1}$$

$$\int_{X \times Y} a(y) d\rho(\mathbf{x}, y) \ < \ +\infty. \tag{2.2}$$

The convexity hypothesis is not restrictive, being satisfied by most loss functions commonly in use. Moreover, it is powerful from a technical point of view since it makes it possible to use convex analysis tools in the study of existence and uniqueness of functional minimizers. Condition 2 is a minimal requirement for defining the expected risk and it is usually satisfied since loss functions commonly in use are continuous. Condition 3 is a technical hypothesis we need in order to use results from convex analysis. Here we note that, for example, it is satisfied in the following cases

1. for the square loss function if

$$\int_{X\times Y} y^2 d\rho(\mathbf{x}, y) < +\infty;$$

   then the condition holds with $p = 2$.

2. If $\ell(y, \cdot)$ is Lipschitz on $\mathbb{R}$ with a Lipschitz constant independent of y and

$$\int_{X\times Y} \ell(y, 0) d\rho(\mathbf{x}, y) < +\infty$$

   then the condition holds with $p = 1$.

According to the above definition exponential loss and the misclassification loss (which simply counts the number of errors) are not admissible.

Remark 2. *We note that the above definition depends on the probability measure ρ. Indeed [CS05] shows that a slightly different definition can be given so that the notion of admissible loss is distribution independent.*

For a loss function ℓ we defined the *expected error* (or *expected risk*) as

$$\mathcal{E}(f) := \mathcal{E}_\rho^\ell(f) = \int_{X\times Y} \ell(y, f(x)) d\rho(x, y).$$

In general we are not ensured that the above integral is finite. The space $\mathcal{F}$ were the integral is finite is sometimes called the target space. For admissible loss function it is straightforward to check that the expected risk, as a functional $\mathcal{E} : L^p(X, \rho_X) \to [0, \infty[$, is well defined. In fact by assumption (2.1)

$$\mathcal{E}(f) \leq \int_{X\times Y} a(y) + b|f(x)|^p d\rho(x, y)$$

for all $f \in \mathcal{F} = L^p(X, \rho_X)$.
If we look at the expected error as a functional on $\mathcal{F}$ we can define the *target function*

$$t_\rho = \operatorname*{argmin}_{f\in\mathcal{F}} \mathcal{E}(f)$$

which is clearly the best possible solution to the learning problem with respect to the chosen loss. The target function can be found constructively considering

$$t_\rho(x) = \operatorname*{argmin}_{w \in \mathbb{R}} \int_Y \ell(y, w) d\rho(y|x).$$

Several target functions can be found in [HTF01].

The choice of the loss function is then crucial as it defines the problem we wish to solve. We note that rather then finding functions which are *similar* to the target (for example similar curves) our main goal is to find some functions whose *expected error* is close to the expected error of the target function. On the other hand we point out that the theory gives the freedom to choose the loss function which is most suited to the problem at hand and it is not always clear if some preferable choice exists. Intuitively the choice should be in principle related to some prior knowledge on the problem. For example, for regression problems it is known that the quadratic loss function is somewhat natural [HTF01] in the presence of Gaussian noise since it arise from a maximum likelihood estimate. Some reasoning along the same line can be found in [PG00] for the ε-insensitive loss function. Moreover it is known that that quadratic costs can be less robust in the presence of points with irregular behavior, that is outliers (see [CS05] and references therein). Actually there are other reasons to consider the quadratic loss function a very natural choice as we will discuss in the next section.
For classification problem one might argue that the most natural choice is the misclassification loss function [Vap98] which simply counts the number of classification errors. Anyway this loss function leads to problem which are intractable from a computational point of view, and convex approximations of it are usually considered [BJM05]. A more principled analysis on the role played by the loss function can be found in [Ste05].

2.2.2.1 Learning with the Square Loss

One reason for which the square loss can be seen as a natural choice for the loss function is that its target function is particularly easy to interpret. In fact it is possible to prove that, if

$$\int_{X \times Y} y^2 d\rho(x, y) \leq \infty$$

the expected error is defined on the space $L^2(X, \rho_X)$ of square integrable functions with respect to ρ_X and the target function is simply the expectation of the conditional probability

$$f_\rho(x) = \int_Y y d\rho(y|x),$$

namely the *regression function*. This can be easily seen writing explicitly the expected error, in fact using the definition of regression function

$$\int_{X\times Y}(y-f(x))^2 d\rho(x,y) = \int_X (f(x)^2 - 2f_\rho(x)f(x))d\rho_X(x) + \int_{X\times Y} y^2 d\rho(y|x)$$

and since the last term is an irreducible error depending on the problem we see that the above quantity is minimized at $f = f_\rho$. Moreover in this case the expected error naturally induces a metric on $\mathcal{F} = L^2(X, \rho_X)$, in fact, for all $f \in L^2(X, \rho_X)$, the following equality is easy to check

$$\mathcal{E}(f) = \|f - f_\rho\|_\rho^2 + \mathcal{E}(f_\rho) \tag{2.3}$$

where $\|f\|_\rho = \int_X f(x)^2 d\rho_X(x)$ is the norm in the $L^2(X, \rho_X)$ and we often refer to it as the ρ-norm. To check the above equation we simply have to add and subtract the regression function in the expression for the expected risk, indeed

$$\begin{aligned}\mathcal{E}(f) &= \int_{X\times Y}(y - f_\rho(x) + f_\rho(x) - f(x))^2 d\rho(x,y) \\ &= \mathcal{E}(f_\rho) + \int_X (f(x) - f_\rho(x))^2 d\rho_X(x) - 2\int_X (y - f_\rho(x))(f(x) - f_\rho(x))d\rho(x,y)\end{aligned}$$

and the last integral is zero by the definition of f_ρ.

Equality (2.3) shows that from a function approximation point of view, the relevant norm in learning (with the square loss) is the ρ-norm since weight is put on the points which are most likely to be sampled. This fact is at the basis of many theoretical studies since we want to ensure some properties (for example convergence) with respect to a metric that we do not know in practice.

2.2.3 Finite Sample Bounds and Consistency

In the first section we argued that a possible way to state the goal of learning is to say that we are looking for an estimator with small expected error. Here we formalize such a statement.
We have seen that each loss function induces a target function which can be seen as the best possible solution to the learning problem since it achieves the minimal error. It is then natural to look at the the deviation $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(t_\rho)$, sometimes called excess error, to measure the generalization property of our estimator. Considering a worst case analysis amounts to study probabilistic inequalities of the form

$$\Pr\left(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(t_\rho) \geq \varepsilon\right) \leq \eta(\varepsilon, n), \tag{2.4}$$

where $\varepsilon > 0$ and $n \in \mathbb{N}$. The above inequality gives a non-asymptotic result since it holds for each fixed number of data n and can be rewritten in various way to better understand its meaning. Assuming we can invert the function $\eta = \eta(\varepsilon, n)$, we can fix a *confidence* $0 < \eta \leq 1$ to get a bound on the excess error

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(t_\rho) \leq \varepsilon(\eta, n),$$

which holds with probability at least $1 - \eta$ for any $n \in \mathbb{N}$. The quantity $1 - \eta$ is often called *confidence*. On the other hand if we fix $\varepsilon > 0$ and $0 < \eta \leq 1$ we have that $n(\varepsilon, \eta)$ is the number of data such that

$$\Pr\left(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(t_\rho) \geq \varepsilon\right) \leq \eta.$$

The last formulation is the one usually considered in the context of PAC (Probably Approximately Correct) learning [Val84] where $n(\varepsilon, \eta)$ is called sample complexity.

One of the reason for studying the excess error is that we are interested to have consistency results. Indeed consistency is a basic property we should require on any learning algorithm.

Definition 2 (Consistency). *We say that an estimator $f_{\mathbf{z}}$ (or the corresponding learning algorithm) is consistent if*

$$\lim_{n\to\infty} \Pr\left(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \geq \varepsilon\right) = 0 \tag{2.5}$$

Moreover we say that $f_{\mathbf{z}}$ is strongly consistent if

$$\Pr\left(\lim_{n\to\infty} \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) = 0\right) = 1$$

If the above convergences hold for all the measure ρ we say they are universal.

Though the above statements are asymptotic they can be seen as a minimal sanity checks ensuring that the learning algorithm performs better as more data are available and eventually leads to the best attainable solution.
Clearly finite sample bounds and consistency are related. If we can prove consistency and convergence rates then we can derive finite sample bounds up-to constant factors. On the other hand finite sample bounds immediately lead to consistency results with convergence rates. For this reason we often talk indifferently of finite sample bounds and convergence rates.

A natural question is whether we can derive consistency and finite sample bounds for a given algorithm for every possible problem, that is uniformly with respect to ρ. Classical results shows that there is a big difference between consistency and finite sample bounds.

In fact it is possible to find algorithms achieving universal consistency, but if we look for convergence rates we have to restrict the class of problems we consider. Intuitively, we can expect that there exists algorithms eventually converging to the target function we might incur into problems such that the convergence rate is arbitrarily slow. The latter results are usually referred to as "no free lunch theorem" [DGL96].

We end this section discussing the connection between the kind of approximation measures considered in Learning Theory and those often considered in related statistical problems. First, another generalization error often considered in statistical learning relates the expected risk to some empirical (and hence controllable) quantity, for example looking at inequalities of the form

$$\Pr\left(|\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})| \geq \varepsilon \right) \leq \eta(\varepsilon, n)$$

where

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_{\mathbf{z}}(x_i))$$

is the empirical error. The analysis of the excess error and empirical error bounds is often related [Bou02]. In our studies we focus on the excess error. From a function approximation perspective we note that looking at the the expected risk is a much weaker requirement than looking for a function $f_{\mathbf{z}}$ which approximates t_ρ point-wise or in some other norm. Clearly in general the expected risk does not even induce a metric on the space of possible solutions. As we previously noted this is the case for the square loss since considering its expected risk corresponds to approximating functions with respect to the norm in $L^2(X, \rho_X)$. In many statistical regression problem the object of interest is the regression function f_ρ so that the least square loss function is implicitly chosen. It is often assumed that f_ρ belongs to some normed space $\mathcal{H}$ (for example Sobolev space) and the approximation is measured with respect to the norm in $\mathcal{H}$. In distribution free non-parametric regression problems approximation of the regression function is measured with respect the norm in $L^2(X, \rho_X)$. Anyway in the latter cases an analysis in expectation is considered.

2.2.4 Hypothesis Space and the Best in the Model

Usually an estimator $f_{\mathbf{z}}$ is selected in a chosen class of possible solutions. We call such a function space the hypothesis space and denote it with $\mathcal{H}$. The choice of the hypothesis space is crucial and has many implications. Again the theory gives the freedom to choose the space $\mathcal{H}$, though some considerations can guide the choice. Indeed its choice should depend on the prior we have on the target function. Let us note some facts related to the choice of the hypothesis space. First of all once an hypothesis space is chosen the best solution is some function whose error is close to

$$\inf_{f \in \mathcal{H}} \mathcal{E}(f). \tag{2.6}$$

Clearly the deviation $\inf_{f\in\mathcal{H}} \mathcal{E}(f) - \mathcal{E}(t_\rho)$ is an approximation error which has nothing to do with the data and is irreducible once the hypothesis space is chosen. This suggests to take an hypothesis space which is "large enough" to make this error negligible, for example if we assume $\mathcal{H}$ to be dense in $\mathcal{F}$ then such an error is zero. Second, in general the above infimum might not be achieved. Existence (as well as uniqueness) depends on the hypothesis space and the loss function. Existence is ensured for example if $\mathcal{H}$ is a Hilbert space and the expected risk $\mathcal{E}(\cdot)$, as a functional from $\mathcal{H}$ to $\mathbb{R}$, is continuous and coercive [1] or if it is just continuous but the space $\mathcal{H}$ is compact. In general though we might choose $\mathcal{H}$ to be a subspace of $\mathcal{F}$ we cannot ensure it to be closed. Uniqueness is ensured if the expected risk is strictly convex [Roc70, ET83a]. If the extremal function of problem (2.6) exists we call it *the best in the model* and denote it with $t_\mathcal{H}$.

We can recognize some different situations:

- the target t_ρ belongs to $\mathcal{H}$. In this case $t_\mathcal{H}$ exists and $t_\mathcal{H} = t_\rho$.
- The target t_ρ does not belong to $\overline{\mathcal{H}}$, which is the closure of $\mathcal{H}$ in $\mathcal{F}$. In this case the best we can aim to is $\inf_{f\in\mathcal{H}} \mathcal{E}(f)$ and $\inf_{f\in\mathcal{H}} \mathcal{E}(f) > \mathcal{E}(t_\rho)$.
- The intermediate case is when the target t_ρ belongs to $\overline{\mathcal{H}}$ (but not to $\mathcal{H}$). In this case even if $t_\mathcal{H}$ does not exist we have $\mathcal{E}(f_\rho) = \inf_{f\in\mathcal{H}} \mathcal{E}(f)$.

To cover all the above three situations we will replace the bound (2.4) with

$$\Pr\left(\mathcal{E}(f_\mathbf{z}) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \geq \varepsilon\right) \leq \eta(\varepsilon, n), \tag{2.7}$$

where $f_\mathbf{z} \in \mathcal{H}$, $\varepsilon > 0$ and $n \in \mathbb{N}$.

2.3 Reproducing Kernel Hilbert Spaces

The so called kernel methods became increasingly popular in recent years because of their generality and their good experimental performance in a variety of different domains. Moreover, one of the reason they are so often advocate is the relationship between kernels and inner product spaces. The use of kernels can be interpreted as an implicit mapping of the input space into a possibly high dimensional *feature* space (via the so called "kernel trick"). From the functional point of view working with kernels amounts to consider hypothesis spaces which are reproducing kernel Hilbert spaces. A reproducing kernel Hilbert

[1] If $(\mathcal{H}, \|\cdot\|_\mathcal{H})$ is a Banach space we say that the functional $\mathcal{E}(\cdot) : \mathcal{H} \to \mathbb{R}$ is coercive (or weakly coercive) if $\lim_{\|f\|\to\infty} \mathcal{E}(f) = \infty$ [Roc70, ET83a].

(RKH) space $\mathcal{H}$ is a Hilbert space of functions which can be completely characterized by a symmetric positive definite function $K : X \times X \to \mathbb{R}$, namely the kernel. From the definition of RKH spaces indeed highlight why they arise naturally in learning problems. In the following we collect the main definition and facts on RKH spaces.

2.3.1 RKH spaces: Definition and main properties

For the moment we let X be a set and $\mathbb{R}^X$ the class of functions $f : X \to \mathbb{R}$.

Definition 3 (RKH space). *A reproducing kernel Hilbert space is an Hilbert space $\mathcal{H} \subset \mathbb{R}^X$ such that for all $x \in X$ there exists a positive constant C_x for which*

$$\|f\|_\infty \leq C_x \|f\|_{\mathcal{H}}, \qquad \forall\ f \in \mathcal{H}. \tag{2.8}$$

An equivalent definition can be given considering the evaluation functional

$$ev_x : \mathcal{H} \to \mathbb{R}, \qquad ev_x(f) = f(x).$$

In fact it is straightforward to check that $\mathcal{H}$ is a RKH space if and only if the evaluation functional is continuous.
Let us some facts. First, $\mathcal{H}$ is a space of *functions* in contrast for example to $L^2(X, \rho_X)$ which are Hilbert spaces of *equivalence classes of functions*. In particular if $f \in \mathcal{H}$ has zero norm then $f(x) = 0$ for all $x \in X$. Second, we note that convergence in the norm $\|\cdot\|_{\mathcal{H}}$ implies point-wise convergence. In fact if $f_i \in \mathcal{H}$ is a sequence converging to some function f in $\mathcal{H}$, then we have

$$|f_i(x) - f(x)| \leq C_x \|f_i - f\|_{\mathcal{H}}$$

by definition of $\mathcal{H}$. Finally we will see in the following that the norm $\|f\|_{\mathcal{H}}$ can often be seen as a functional measuring the regularity of f and this becomes important in the definition of several learning algorithms.

A very important fact is that to each RKH space is naturally associated a (unique) positive definite kernel called reproducing kernel (see example in table 2.1). We recall the following definition.

Definition 4. *A positive definite (PD) kernel is a symmetric function $K : X \times X \to \mathbb{R}$ such that, for all $N \in \mathbb{N}$, $x_1, \ldots, x_N \in X$ and $c_1, \ldots, c_N \in \mathbb{R}$, we have*

$$\sum_{i,j=1}^{N} c_i c_j K(x_i, x_j) \geq 0.$$

The relationship between RKH spaces and PD kernels is straightforward recalling that Rietz theorem [Lan93] ensures the existence of a unique element $K_x \in \mathcal{H}$ such that

$$ev_x(x) = \langle f, K_x \rangle_{\mathcal{H}}, \quad \forall\ f \in \mathcal{H}$$

and we can easily prove that $K(x,s) = \langle K_x, K_s \rangle_{\mathcal{H}}$ is a PD kernel. Clearly, by definition of reproducing kernel, we have the reproducing property

$$f(x) = \langle f, K_x \rangle .$$

The following important result shows that we can invert the relationship between RKH spaces and PD kernel.

Theorem 1. *Let $K : X \times X \to \mathbb{R}$ be a positive definite kernel. Then there exists a unique RKH space $\mathcal{H} \subset \mathbb{R}^X$ with K reproducing kernel.*

The proof relies on showing that if we let $K_x = K(x, \cdot)$, the space $\mathcal{H}$ induced by the kernel K can be built as the completion of the finite linear combinations $f = \sum_{i=1}^{N} c_i K_{x_i}$ with respect to the inner product $\langle K_s, K_x \rangle_{\mathcal{H}} = K(s,x)$.
Most properties of functions in a RKH space can be described in terms of properties of the associated reproducing kernel. In what follows we do not want to give a complete characterization of the relation between properties of elements in $\mathcal{H}$ and properties of the corresponding reproducing kernel and we describe some results in the specific setting that we will consider throughout. In particular we will consider the following:

Assumption 1. *We assume that the kernel K is continuous and bounded, i.e. there exists a positive constant κ such that*

$$\sup_{x \in X} \sqrt{K(x,x)} \leq \kappa.$$

It is easy to note that, if the kernel is bounded we can take $C_x \leq \kappa$ in (2.8), in fact by Schwartz inequality

$$|f(x)| \leq \langle K_x, f \rangle_{\mathcal{H}} \leq \|K_x\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \leq \kappa \|f\|_{\mathcal{H}} \tag{2.9}$$

where we used $\|K_x\|_{\mathcal{H}} = \sqrt{\langle K_x, K_x \rangle_{\mathcal{H}}} = \sqrt{K(x,x)}$. In this case it is straightforward to see that convergence in $\mathcal{H}$ implies uniform convergence.

The following proposition highlights how properties of the kernel induce properties of the elements in $\mathcal{H}$.

Proposition 1. *We let X be a Polish space endowed with a probability measure ρ_X and $\mathcal{H}$ a RKH space with kernel K. If assumption 1 holds then we have that*

- *the elements of $\mathcal{H}$ are continuous (and hence measurable),*
- *the elements of $\mathcal{H}$ belong to $L^p(X, \rho_X)$.*

name	expression	kernel parameters
linear	$K(x,s) = x \cdot s$	
polynomial	$K(x,s) = (x \cdot s)^p$	p
gaussian	$K(x,s) = e^{-\frac{\|x-s\|^2}{2\sigma^2}}$	σ

Table 2.1: Some examples of classical kernel functions.

2.3.2 Operators Defined by a Kernel

We now define some operators which will be useful in the following (see [CDVT05] for details). Let us first recall some notation and definitions if $\mathcal{H}$ and $\mathcal{G}$ are two Hilbert spaces we denote with $\mathcal{B}(\mathcal{H},\mathcal{G})$ the Banach space of bounded operators endowed with the operator norm $\|\cdot\|_{\mathcal{B}(\mathcal{H},K)}$. When it is clear we simply denote the operator norm with $\|\cdot\|$ and write $\mathcal{B}(\mathcal{H})$ if $\mathcal{G} = \mathcal{H}$. Moreover let $\mathcal{B}_2(\mathcal{H},K)$ be the space of Hilbert-Schmidt operators, which is an Hilbert space of operators A such that $\mathrm{Tr}(A^*A) \leq \infty$ endowed with the scalar product $\langle A,B\rangle_2 = \mathrm{Tr}\, A^*B$. If $|A| = (A^*A)^{1/2}$, let $\mathcal{B}_1(\mathcal{H},\mathcal{G})$ be the space of trace class operators such that $\mathrm{Tr}(|A|) \leq \infty$ endowed with the trace norm $\|A\|_1 = \mathrm{Tr}(|A|)$. Finally we denote with $Im(A)$ the image of A, with $ker(A)$ its kernel and with A^* the adjoint of A.

We first introduce the inclusion operator $I_K : \mathcal{H} \to L^2(X,\rho_X)$, whose explicit form is

$$I_K f(x) = \langle f, K_x\rangle_{\mathcal{H}}, \qquad x \in X.$$

Such an operator is continuous because of assumption 1. Moreover we consider the following operators: the adjoint operator $I_K^* : L^2(X,\rho_X) \to \mathcal{H}$, the covariance operator $T : \mathcal{H} \to \mathcal{H}$ defined by $T = I_K^* I_K$ and the operator $L_K : L^2(X,\rho_X) \to L^2(X,\rho_X)$ defined by $L_K = I_K I_K^*$. It can be easily proved that

$$\begin{aligned} I_K^* f &= \int_X K_x f d\rho_X(x), \qquad f \in L^2(X,\rho_X) \\ T &= \int_X \langle \cdot, K_x\rangle_{\mathcal{H}} K_x d\rho_X(x), \\ L_K f &= \int_X K(x,\cdot) f d\rho_X(x), \qquad f \in L^2(X,\rho_X). \end{aligned}$$

Since the kernel is bounded and positive definite, both L_K and T are trace class positive operator and there is a sequence of vectors $(v_i)_{i\geq 1}$ in $\mathcal{H}$ and a sequence of numbers $(\sigma_i)_{i\geq 1}$ (possibly finite) such that

$$Tf = \sum_{i=1} \sigma_i \langle f, v_i\rangle_{\mathcal{H}} v_i, \qquad \langle v_i, v_j\rangle_{\mathcal{H}} = \delta_{ij}, \qquad \sum_i \sigma_i \leq \kappa^2, \qquad \sigma_{i+1} \geq \cdots \geq \sigma_i > 0,$$

for all $f \in \mathcal{H}$. Letting $u_i = \frac{1}{\sqrt{\sigma_i}} v_i \in L^2(X, \rho_X)$

$$L_K f = \sum_{i=1}^{n} \sigma_i \langle f, u_i \rangle_\rho u_i \qquad \langle u_i, u_j \rangle_\rho = \delta_{ij}.$$

In particular, $\|L_K\|_{\mathcal{B}(L^2(X,\rho_X))} = \|T\|_{\mathcal{B}(\mathcal{H})} \leq \sum_i \sigma_i \leq \kappa^2$.

For $f \in \mathcal{H}$ we can relate the norm in $\mathcal{H}$ and $L^2(X, \rho_X)$ using T. In fact if we regard $f \in \mathcal{H}$ as a function in $L^2(X, \rho_X)$ we can write

$$\|f\|_\rho = \left\| \sqrt{T} f \right\|_{\mathcal{H}}. \tag{2.10}$$

This fact can be proved recalling that since the inclusion operator is continuous it admits a polar decomposition $I_K = U\sqrt{T}$, where U is a partial isometry [Rud91].

Replacing ρ_X by the empirical measure $\rho_{\mathbf{x}} = 1/n \sum_{i=1}^{n} \delta_{x_i}$ on a sample $\mathbf{x} = (x_1, \ldots, x_n)$ we can define the empirical counterpart of the above operators. The sampling operator $S_{\mathbf{x}} : \mathcal{H} \to \mathbb{R}^n$ is defined by $(S_{\mathbf{x}} f)_i = f(x_i) = \langle f, K_{x_i} \rangle_{\mathcal{H}}$; $i = 1, \ldots, n$, where the norm $\|\cdot\|_n$ in $\mathbb{R}^n$ is $1/n$ times the euclidean norm. Moreover we can define the adjoint $S_{\mathbf{x}}^* : \mathbb{R}^n \to \mathcal{H}$, the empirical covariance operator $T_{\mathbf{x}} : \mathcal{H} \to \mathcal{H}$ such that $T_{\mathbf{x}} = S_{\mathbf{x}}^* S_{\mathbf{x}}$ and the operator $S_{\mathbf{x}} S_{\mathbf{x}}^* : \mathbb{R}^n \to \mathbb{R}^n$. It follows that for $\xi = (\xi_1, \ldots, \xi_n)$

$$\begin{aligned} S_{\mathbf{x}}^* \xi &= \frac{1}{n} \sum_{i=1}^{n} K_{x_i} \xi_i, \\ T_{\mathbf{x}} &= \frac{1}{n} \sum_{i=1}^{n} \langle \cdot, K_{x_i} \rangle_{\mathcal{H}} K_{x_i}, \\ S_{\mathbf{x}} S_{\mathbf{x}}^* &= \frac{1}{n} \mathbf{K} \end{aligned}$$

where $\mathbf{K}$ is the kernel matrix such that $(\mathbf{K})_{ij} = K(x_i, x_j)$.

2.3.3 Feature Map

At the beginning of this section we motivated the use of kernels for their property of implicitly mapping the input data in a higher dimensional space. We end our brief introduction to RKH spaces describing this last point of view.

We let $\hat{\mathcal{H}} \subset \mathbb{R}^X$ be an Hilbert space of functions. Consider a map $\Phi : X \to \hat{\mathcal{H}}$ and $K(x, s)$ to be some real valued function on $X \times X$. The condition

$$K(x, s) = \langle \Phi(x), \Phi(s) \rangle_{\hat{\mathcal{H}}} \tag{2.11}$$

for $x, s \in X$, is equivalent to

- the function K is a reproducing kernel.
- If we let $A : \hat{\mathcal{H}} \to \mathbb{R}^X$ be the map defined as

$$Af(x) = \langle \Phi(x), f \rangle_{\hat{\mathcal{H}}}$$

  then A is a partial isometry from $\hat{\mathcal{H}}$ onto the RKH space $\mathcal{H} \subset \mathbb{R}^X$ with kernel K.

If equation (2.11) holds we call Φ a feature map.

Some examples of feature maps are the following.

- Let $\Phi(x) = K_x$, in this case $\hat{\mathcal{H}} = \mathcal{H}$.
- Let $(e_\gamma)_{\gamma \in \Gamma}$ be an orthonormal basis in $\mathcal{H}$ then we can choose $\Phi(x) = (e_\gamma(x))$. In this case $\hat{\mathcal{H}} = \ell^2$, which is the space of square summable sequences, and $K(x, s) = \sum_{\gamma \in \Gamma} e_i(x) e_i(s)$.

2.4 Empirical Risk minimization

In this section we focus on a specific learning algorithm, the Empirical Risk Minimization (ERM). The importance of ERM lies in the fact the most algorithms can be seen as refinements of it. In a few words the idea is that since we cannot minimize the expected error directly we can replace it with its empirical counterpart. Usually the minimization is restricted to a certain class of candidate estimators of the regression (or the target function), that is the hypothesis space.

The question is then what kind of performance we can expect from this algorithm and what are the main factors affecting it. The intuition suggests that the following aspects should be relevant:

- the number of available examples n. The more examples we have the more we expect the empirical error to be a good estimator of the expected error. In general we cannot assume to have "*enough*" examples.
- The hypothesis space $\mathcal{H}$. On this respect we can note some interesting facts. We noticed that once we choose the hypothesis space we introduce a *bias* in the problem, meaning that we cannot aim at functions which are outside of $\mathcal{H}$. Then we might say that we would like $\mathcal{H}$ to be as "big" as possible. On the other hand simple examples suggest that if we have complex enough hypothesis space we might incur into over-fitting. Since the number of available data is limited we risk to over-estimate the

information at hand loosing generalization properties. Then it looks like we should look for some intermediate "size" for the hypothesis space depending on the available data.

In this section we give a formal and quantitative derivation of the above intuitions.

The analysis of ERM received a lot of attention and a great number of results exists. Here we do not aim neither at giving an exhaustive account of the literature nor at following the historical derivation of the analysis of ERM. We choose to present an approach proposed in [CS02b] for the quadratic loss and extended in [RDVC+04] to a wider class of loss functions. Such an approach relies on an interesting functional analytical framework which is at the basis of the developments in the next chapters.

2.4.1 Consistency of ERM on Kernel Classes

The ERM algorithm defines an estimator which minimizes the error on the training set for a given loss function so that

$$f_{\mathbf{z}} = \operatorname*{argmin}_{f \in \mathcal{H}} \mathcal{E}_{\mathbf{z}}(f)$$

where

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

is the empirical error. In general existence of the above minimizer is not ensured but it is always possible to consider functions which are ε-close to the infimum, namely ε-minimizers (see [MNPR04] and references therein). For sake of simplicity we assume that $f_{\mathbf{z}}$ exists. Throughout this section we assume the following conditions hold.

1. The output space is a bounded subset

$$Y = [-M, M] \subset \mathbb{R} \tag{2.12}$$

   for some $M > 0$. Moreover the input space X is compact.

2. The hypothesis space is a *ball* in a RKH space that is we consider

$$\mathcal{H}_R = \{f \in \mathcal{H} : \ \|f\|_{\mathcal{H}} \leq R\}$$

   sometimes called a kernel class. In this case it is clear that the functions in $\mathcal{H}_R$ are bounded, in fact it follows from (2.9) that $\|f\|_\infty \leq \kappa R$. Moreover as we choose to work in the space $\mathcal{H}_R$ the best we can aim at is the function f^R which solves the problem

$$\min_{f \in \mathcal{H}_R} \mathcal{E}(f)$$

problem	loss	L	C_0
regr	quad	$2\kappa R + M$	M^2
regr	abs val	1	M
regr	ε-insensitive	1	M
class	quad	$2\kappa R + 2$	1
class	hinge	1	1
class	logistic	$e^{\kappa R}/(1 + e^{\kappa R})$	$(\ln 2)$

Table 2.2: Values of L and C_0 for a number of loss functions for regression (regr) and classification (class).

so that we look for bounds on $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f^R)$.

3. The loss function is admissible. This implies in particular that [Roc70, ET83a]

   - The loss function is a Lipschitz function on each bounded interval, i.e. for every $I \subset \mathbb{R}$ there exists a constant $L = L_I$, which depends on I, such that

     $$|\ell(y, w_1) - \ell(y, w_2)| \leq L|w_1 - w_2|, \qquad w_1, w_1 \in I. \tag{2.13}$$

     In particular if we consider functions in $\mathcal{H}_R$ we can take $I = [-\kappa R, \kappa R]$.
   - There exists a constant C_0 such that, $\forall y \in Y$,

     $$\ell(y, 0) \leq C_0. \tag{2.14}$$

The explicit values of L and C_0 depend on the specific form of the loss function as can be seen in table 2.2.

2.4.2 Error Bounds and Covering Numbers

The idea to study the behavior of the ERM algorithm is to consider the error decomposition

$$|\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f^R)| \leq |(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}))| + |(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f^R))| + |(\mathcal{E}_{\mathbf{z}}(f^R) - \mathcal{E}(f^R))|$$

where the middle term is always not positive by definition of $f_{\mathbf{z}}$. A possible approach relies on controlling the deviation of the empirical risk to the expected risk for all the functions in $\mathcal{H}_R$, that is we look for bounds on

$$\sup_{f \in \mathcal{H}_R} |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)|. \tag{2.15}$$

Clearly

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f^R) \leq 2 \sup_{f \in \mathcal{H}_R} |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)|.$$

The study of the tails of the stochastic process in (2.15) has been studied extensively in the theory of probability (see for example [vdVW96] or [Bou02] and references therein) and can be seen as a formulation of the law of large numbers which is uniform for a given function space. To derive this kind of probabilistic results we need a quantitative description of the capacity (complexity) of the space $\mathcal{H}_R$. To this end we recall that for a compact set in a metric space the covering number $\mathcal{N}(S, a)$ is defined as the minimal integer $m \in \mathbb{N}$ such that there exist m balls $B_i(a),\ i = 1, \ldots, m$ with radius a covering S, that is $S \subset \cup_{i=1}^m B_i(a)$. Since $\mathcal{H}_R$ is a compact subset of $\mathcal{C}(X)$, the space of continuous functions on X, we can define the covering number $\mathcal{N}(R, a)$ of $\mathcal{H}_R$ with respect to norm $\|\cdot\|_\infty$. The following result is a rather straightforward extension of theorem C in [CS02b].

Theorem 2. *If conditions (2.12), (2.13) and (2.14) hold, then for all $\varepsilon > 0$ and $n \in \mathbb{N}$ the following inequality holds*

$$\Pr\left(\sup_{f \in \mathcal{H}_R} |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)| \geq \varepsilon \right) \leq 2\mathcal{N}(R, \frac{\varepsilon}{4L}) e^{-\frac{n\varepsilon^2}{8(L\kappa R + C_0)^2}}.$$

We note that the above approach is not restricted to the case of RKH spaces and extend to any space $\mathcal{H}$ which can be compactly embedded in $\mathcal{C}(X)$. We postpone the proof of the above theorem to see how it immediately yields finite sample bounds for the ERM algorithm.

Corollary 1. *If conditions (2.12), (2.13) and (2.14) hold then for all $0 < \eta \leq 1$ and $n \in \mathbb{N}$ the following inequality holds*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f^R) \leq 2\varepsilon(\eta, n, R),$$

and moreover

$$\mathcal{E}(f_{\mathbf{z}}) \leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \varepsilon(\eta, n, R). \tag{2.16}$$

$\varepsilon(\eta, n, R)$ is obtained inverting

$$\eta(\varepsilon, n, R) = 2\mathcal{N}(R, \frac{\varepsilon}{4L}) e^{-\frac{n\varepsilon^2}{8(L\kappa R + C_0)^2}} \tag{2.17}$$

with respect to ε.

To get a more explicit version of the above bounds we need an explicit estimate of the covering number (see [Zho03] and references therein). For example we recall that if $X \subset \mathbb{R}^d$

and the kernel K is a $\mathcal{C}^\infty$ function then the corresponding $\mathcal{H}$ can be embedded into the Sobolev space[2] H^s for any $s \in \mathbb{N}$. In this case it is known (see for example [Zho03]) that

$$\ln \mathcal{N}(R, a) \leq \left(\frac{C_h R}{a} \right)^{\frac{2d}{h}}.$$

In Learning Theory other notions of capacity of a function space have been considered, we name, for example, various notions of empirical covering numbers and Rademacher or Gaussian complexities (see [Bou02] and references therein). Indeed a lot of studies focused on deriving tight concentration results for the supremum of the empirical process in (2.15). Moreover another source of improvement relies on the observation that looking for uniform results over all $\mathcal{H}_R$, when we might actually be searching just a small subspace of $\mathcal{H}_R$, can be a loose approach. This is at the basis of recent developments, the so called localized approaches (see again [Bou02] and references therein). Anyway it should be noted that classical results [Vap98, ABDCBH97] show that uniform convergence is not only sufficient but also necessary to capture the property of *universal* consistency for ERM.

An interesting aspect of the above analysis can derived from inequality (2.16). The first term at the left hand side is clearly decreasing in R while the complexity term $\varepsilon(\eta, n, R)$ is increasing (since the covering numbers increase with R). This suggests that rather then just minimizing the empirical error we should try to simultaneously control the complexity of the solution. The latter point inspired the so called Structural Risk Minimization principle (SRM) [Vap98]. Here we discuss a particular instance of such a principle.

2.4.3 Proof

We now give the proof of theorem 2.

Proof 1. *We let*

$$\Delta_{\mathbf{z}}(f) := \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f),$$

and it is straightforward to check that inequality (2.13) implies

$$|\Delta_{\mathbf{z}}(f_1) - \Delta_{\mathbf{z}}(f_2)| \leq 2L \|f_1 - f_2\|_\infty, \quad \forall f_1, f_2 \in \mathcal{H}_R. \tag{2.18}$$

Moreover we let $(f_i)_{i=1}^m$ the center radius of the balls with radius b covering $\mathcal{H}_R$. Then $\forall f \in \mathcal{H}_R$ there exists a center f_i such that

$$|\Delta_{\mathbf{z}}(f)| = |\Delta_{\mathbf{z}}(f) - \Delta_{\mathbf{z}}(f_i) + \Delta_{\mathbf{z}}(f_i)| \leq 2Lb + |\Delta_{\mathbf{z}}(f_i)|$$

[2]Here H^h is the space of the square integrable functions such that the distributional derivative $D^\alpha f$ is square integrable for all $(\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}_+^d$ with $\alpha_1 + \cdots + \alpha_d \leq s$.

where we used the triangle inequality, inequality (2.18) and the definition of covering. The above inequality suggests that if we take $b = \varepsilon/4L$ and we can ensure that $\Delta_{\mathbf{z}}(f) \leq \varepsilon/2$ with high probability then $|\Delta_{\mathbf{z}}(f)| \leq \varepsilon$ uniformly over $\mathcal{H}_R$ with high probability. In fact if we let $m = \mathcal{N}(R, \frac{\varepsilon}{4L})$ from elementary probability we get

$$\Pr\left(\sup_{\mathcal{H}_R} |\Delta_{\mathbf{z}}(f)| \geq \varepsilon\right) \leq \Pr\left(\cup_{i=1}^{m} |\Delta_{\mathbf{z}}(f_i)| \geq \frac{\varepsilon}{2}\right) \leq$$
$$\sum_{i=1}^{m} \Pr\left(|\Delta_{\mathbf{z}}(f_i)| \geq \frac{\varepsilon}{2}\right) = \mathcal{N}(R, \frac{\varepsilon}{4L}) \Pr\left(|\Delta_{\mathbf{z}}(f)| \geq \frac{\varepsilon}{2}\right). \tag{2.19}$$

For fixed f we can control the last term in the above chain of inequalities defining the random variable $\xi = \xi(x, y) = \ell(y, f(x))$. In fact for all $f \in \mathcal{H}_R$ triangle inequality, inequality (2.13) and (2.14) ensure

$$|\ell(y, f(x))| \leq |\ell(y, f(x) - \ell(y, 0)| + |\ell(y, 0)| \leq L \|f\|_\infty + C_0 \leq LR\kappa + C_0$$

so that ξ is bounded (and positive by definition of loss function). Moreover

$$\frac{1}{n} \sum_{i=1}^{n} \xi_i = \mathcal{E}_{\mathbf{z}}(f), \qquad \mathrm{E}[\xi] = \mathcal{E}(f),$$

so that a direct application of Höeffding inequality leads to

$$\Pr\left(|\Delta_{\mathbf{z}}(f)| \geq \frac{\varepsilon}{2}\right) \leq 2e^{-\frac{n\varepsilon^2}{8(L\kappa R + C_0)^2}}$$

The theorem is proved plugging the above inequality into (2.19).

2.5 From Ivanov to Tikhonov Regularization

In our setting we can consider the one parameter family of estimators $f_{\mathbf{z}}^R$ obtained minimizing the expected risk for different values of the radius $R > 0$. Let us look in some more detail at this modified algorithm. Indeed the crucial problem is how to choose a value R_n for the radius, as a function of the available data and possibly depending on the prior we have on t_ρ, in such a way that $\mathcal{E}(f_{\mathbf{z}}^{R_n}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)$ is small with high probability. The analysis previously done suggests a possible way to choose R. In fact if we can find a bound on the approximation term,

$$\mathcal{E}(f^R) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) \leq A(R),$$

then for any $R > 0$ we have

$$\mathcal{E}(f_{\mathbf{z}}^R) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \leq \mathcal{E}(f_{\mathbf{z}}^R) - \mathcal{E}(f^R) + \mathcal{E}(f^R) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \leq 2\varepsilon(\eta, n, R) + A(R)$$

and we can choose the value R_n optimizing the above bound. In the above decomposition we recognize a bias-variance problem. The first term is due to the random sampling and is sometimes called sample error, whereas the second term is independent to the data and is called approximation error. Unfortunately, as we previously mentioned, to find the bound on the approximation term we need to put some condition on t_ρ. It is possible to see that in this way the proposed choice for R depends on regularity assumptions which are usually unknown in practice. For example we might assume that the target is in some Sobolev space but we might not know the smoothness index of the space. More specifically for the square loss it can be shown [CS02b, SZ03] that if

$$f_\rho \in \{f \in L^2(X, \rho_X) : f = L_K^r w, \|w\|_\rho \leq R^*, w \in L^2(X, \rho_X)\}$$

then we can take $A(R) \propto R^{-r}$. Again in practice the correct value of the index r is unknown. Such parameter choice is usually called *a priori*, it depends on the number of examples $R = R(n)$ but *not* on the sample $\mathbf{z}$. Clearly it would be much more interesting to have an a posteriori, data driven, parameter choice $R = R(n, \mathbf{z})$ independent to the regularity of t_ρ. The principle of SRM suggests that a possible choice for R is the value minimizing the r.h.s. of (2.16). Such approach is also called complexity regularization [DGL96] and was proved to be adaptive to the unknown regularity of t_ρ in a context similar to the one considered here. The concept of *adaptivity* is extremely important in fact it means that we can choose the parameter R without knowing the regularity of the target but we achieve the same performances as if we knew it in advance.

A main drawback in the above algorithm is that its implementation is indeed cumbersome. For each value of R we should solve a different minimization problem. If we write the problem as

$$\begin{aligned} &\min_{f\in\mathcal{H}} \mathcal{E}_{\mathbf{z}}(f) \\ &s.t. \|f\|_{\mathcal{H}}^2 \leq R \end{aligned} \tag{2.20}$$

it is tempting to consider the Lagrangian formulation of the above constrained minimization problem taking

$$\min_{f\in\mathcal{H}, g\in\mathcal{B}} \{\mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|_{\mathcal{H}}^2\}. \tag{2.21}$$

The latter class of algorithms is well-known and it is called kernel ridge regression in statistics [HTF01] and regularization networks in machine learning [EPP00]. The role played by R is now played by λ.

One of the main motivation for our studies was the observation done by many authors that the described algorithms are essentially a generalization of the so called Ivanonv and Tikhonov regularization algorithms that are a well known algorithms for the solution of ill-posed inverse problems. In inverse problems the goal is to find stable approximation to a possibly ill-posed problem and this is usually done restricting the space of possible solution. This analogy which has been noted but never investigated in depth seems interesting for various reasons. The theory of regularization of ill-posed problems is a well established theory and a lot of results are available. In particular many different regularization algorithms exist, besides the two we presented above, and they are well studied both from the experimental and the theoretical point of view. Moreover one of the central problems in regularization theory is to select the free parameters such as R or λ which are called regularization parameters. Drawing a clear mathematical connection between the theory of inverse problems and learning theory would hopefully allows to transfer results and algorithms from one theory the other and eventually increase the understanding of both. The crucial question is then if learning can be written as an inverse problem. In the next chapter we will give a possible answer to this question.

We end this section with the following remark on a tricky difference between deterministic problems and probabilistic problems when considering the relation between Lagrangian and constrained formulation of a minimization problem.

Remark 3 (Ivanov vs Tikhonov). *Once we pass from problem (2.20) to problem (2.21) a very natural question is how to choose λ. A reasonable guess is that if we know how to choose R then we can immediately have a choice for λ. A closer inspection shows that this is not so straightforward. Indeed in a deterministic setting the above two formulations of the minimization problem are equivalent, meaning that it if we let $f_{\mathbf{z}}^{\lambda}$ be the minimizer of (7) there exists a value λ^* such $f_{\mathbf{z}}^{\lambda^*}$ is a solution of (2.20). Similarly we can pass from (2.20) to (7). Unfortunately probability makes things tricky. For example to pass from R to λ we should find λ such that*

$$R = \left\| f_{\mathbf{z}}^{\lambda} \right\|_{\mathcal{H}}.$$

Since $f_{\mathbf{z}}^{\lambda}$ depends on the training set $\mathbf{z}$ the above relation is no longer deterministic and deserves some more analysis. A possible way out of this is noting that it is often the case that (see end of chapter 5)

$$\left\| f_{\mathbf{z}}^{\lambda} \right\|_{\mathcal{H}} \leq \frac{C}{\sqrt{\lambda}}$$

for some constant $C > 0$ and we can take $R = \frac{C}{\sqrt{\lambda}}$. Anyway in general such approach is loose and give suboptimal results. Moreover we somewhat forgot about the explicit form of $f_{\mathbf{z}}^{\lambda}$ to rewrite an algorithm similar to ERM.

Chapter 3

Learning, Regularization and Inverse Problems

We have seen that the analysis of ERM suggests that the key to obtain a meaningful solutions to learning problems is to control the complexity of the hypothesis space. On the other hand in ill-posed inverse problems usually is the notion of smoothness to play a main role. In both domains it is crucial to restrict the space of possible solutions. Not surprisingly the form of the algorithms proposed in both theories is strikingly similar [MRP02] and the point of view of regularization is indeed not new to learning [PG92, EPP00, Vap98, Arb95, Fin99, Kec01, SS02, MNPR04, PRMN04].

Anyway a closer look shows that a rigorous mathematical connection between learning theory and the theory of ill-posed inverse problems is not straightforward since the settings underlying the two theories are different.

To investigate such a connection we restrict the focus on the quadratic loss function and hypothesis spaces which are RKH spaces so that we can cast the problem of learning in a functional analytical framework which is ideal to highlight the connections with the theory of inverse problems.

Our analysis and contribution develop in two steps. First, we rewrite the problem of learning as an inverse problem. Indeed we show that the problem of finding the best in the model can be translated in the problem of inverting a linear embedding equation. As we consider the problem on the data we see that this is nothing but a stochastic discretization of the aforementioned continuous problem. This clarifies in particular the following important fact. The sampling induces a linear problem that can be seen as a perturbation (due to random discretization) of the linear inverse problem corresponding to the population case. Regularized solutions in learning should not only provide stable approximate solutions to the unstable, discrete problem but especially give approximations of the solution to the ill-

posed infinite dimensional problem. Unlike the inverse problem setting, in learning *stability* is meant with respect to perturbations on the problem due to the random sampling. This last fact requires the definition of different notions of perturbation measures compared to the one usually considered in inverse problems.
Second, building up on the connection between learning and inverse problems we investigate in a systematic way the notion of regularization for learning with the square loss. In the context of learning the term regularization usually refers to techniques to avoid overfitting. Typically, regularization boils down to a Lagrangian formulation of an appropriate constrained minimization problem - e.g. Tikhonov regularization. On the other hand in the context of inverse problems regularization is defined in an abstract way by means of a set of simple conditions. Our goal is to discuss how this definition extends to learning. The outcome of our studies is essentially that the same definition of regularization holds for inverse problems and learning. We can see some differences when the best in the model does not exist since this situation is usually not considered in inverse problems. As we discuss in the next chapter, in this case we need an extra condition to perform the required probabilistic analysis. In any case we can analyze in a unified framework a large class of regularization schemes used in the context of inverse problems and show that they give rise to consistent kernel methods. Several new algorithms are presented and old ones are revisited. From the theoretical point of view a deeper understanding on the conditions under which the above algorithms work can be given discussing which are the suitable prior assumptions on the problem. Again it turns out that the connection with inverse problems is fruitful. The notion of prior often considered in learning is a special case kind of source condition for ill-posed problems. Indeed following what is standard in inverse problems we can consider more general priors. Finally an intuitive explanation of the way such algorithms work can be given from a filter function point of view.

The chapter is divided as follows. In section 3.1 we give an overview of the main concepts in the of ill-posed inverse problems. In section 3.2 we discuss the connection between learning and inverse problems. In section 3.3 we give an abstract characterization of regularization. In section 3.4 we discuss several examples of algorithms satisfying the aforementioned definition. In section 3.5 we propose an interpretation of regularization from a filter function perspective. In section 3.6 we describe a class of a priori assumption for approximation schemes in reproducing kernel Hilbert spaces. Finally in section 3.7 we summarize our results and discuss some connections with previous works.

3.1 Ill-Posed Inverse Problems and Regularization

In this section we give a very brief account of the main concepts of linear inverse problems and regularization theory (see [TA77, Gro84, BDMP85, BDMP88, EHN96, TGSY95] and references therein).

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces and $A : \mathcal{H} \to \mathcal{G}$ a linear bounded operator. Consider the equation

$$Af = g \tag{3.1}$$

where $g \in \mathcal{G}$ is the *exact* datum. Finding the function f satisfying the above equation, given A and g, is the linear inverse problem associated to (3.1). In general the above problem is *ill-posed*, that is, the solution either do not exists ($Im(A) \subset \mathcal{G}$) , is not unique ($\ker(A) \neq \emptyset$) or does not depend continuously on the datum g (the inverse of A is not continuous). Existence and uniqueness can be restored introducing the Moore-Penrose (or generalized) solution $f^\dagger$ defined as the minimal norm solution of the least squares problem [1]

$$\min_{f \in \mathcal{H}} \|Af - g\|_{\mathcal{G}}^2 . \tag{3.2}$$

It can be shown [TGSY95] that the generalized solution $f^\dagger$ exists if and only if $Pg \in Im(A)$, where $P : \mathcal{G} \to \overline{Im(A)}$ is the projection on the closure of the range of A. If we differentiate (3.2) we immediately see that $f^\dagger$ solves the normal equation

$$A^*Af = A^*g.$$

Recall that if $Im(A)$ is not closed the inverse operator $(A^*A)^{-1}$ is unbounded (in fact we should consider the generalized inverse $A^\dagger g = f^\dagger$) it is then clear that the generalized solution $f^\dagger$ does not depend continuously on the datum g, so that finding $f^\dagger$ is again an ill-posed problem. This is not a trivial problem since the exact datum g is usually not known and only a *noisy* datum $g_\delta \in \mathcal{G}$ is given. In the simplest deterministic model $\|g - g_\delta\|_{\mathcal{G}} \leq \delta$ where, $\delta > 0$ is the noise level [2]. This last fact can be better understood considering the case when the operator A is compact. In this case, using the singular system (σ_i, u_i, v_i) associated to A (see Appendix), we can write the generalized solution explicitly as

$$f^\dagger = \sum_{i=1}^{\infty} \frac{1}{\sigma_i} \langle g, u_i \rangle_{\mathcal{G}} \, v_i.$$

Even if we have small perturbations on the datum g they might reflect in oscillatory behavior of $f^\dagger$ in correspondence of small singular values.

Regularization theory is developed to obtain a stable solution to ill-posed problems. For example according to Tikhonov regularization [TA77] a possible way to find a solution depending continuously on the data is to replace Problem (3.2) with the following convex problem

$$\min_{f \in \mathcal{H}} \{ \|Af - g_\delta\|_{\mathcal{G}}^2 + \lambda \|f\|_{\mathcal{H}}^2 \}. \tag{3.3}$$

[1] It is easy to prove [EHN96] that the set of all possible solutions of Problem (3.2) is closed and convex so that a unique element of minimal norm exist.

[2] The framework we described can be easily extended to the case of a noisy operator $A_\delta : \mathcal{H} \to \mathcal{G}$ where $\|A - A_\delta\| \leq \delta$, and δ is called noise on the model [TGSY95].

Indeed the term $\|f\|^2_{\mathcal{H}}$ enforces a smooth behavior and hence stability. In fact for $\lambda > 0$, the unique minimizer of the above problem is given by

$$f^\lambda_\delta = (A^*A + \lambda I)^{-1} A^* g_\delta. \tag{3.4}$$

We can see that Tikhonov regularization replaces operator $(A^*A)^{-1}$ (or $A^\dagger$) with the a one parameter family of operators $(A^*A + \lambda I)^{-1}$, namely the regularization operators. If A is compact we can have a better intuition of the way the above algorithm works since we can write

$$f^\lambda_\delta = \sum_{i=1}^{\infty} \frac{\sigma}{\sigma_i^2 + \lambda} \langle g_\delta, u_i \rangle_{\mathcal{G}} \, v_i$$

and we see that if we choose λ correctly we can control the behavior of the solution in correspondence of small singular values. The crucial issue is then the choice of the regularization parameter λ which should be chosen in such a way that the *reconstruction error* $\left\| f^\lambda_\delta - f^\dagger \right\|_{\mathcal{H}}$ is small. In particular, λ must be selected, as a function of the noise level δ and the data g_δ, in such a way that the regularized solution $f^{\lambda(\delta, g_\delta)}_\delta$ converges to the generalized solution, that is,

$$\lim_{\delta \to 0} \left\| f^{\lambda(\delta, g_\delta)}_\delta - f^\dagger \right\|_{\mathcal{H}} = 0, \quad \text{and} \quad \lim_{\delta \to 0} \lambda(\delta, g_\delta) = 0, \tag{3.5}$$

for any g such that $f^\dagger$ exists. The one parameter family of regularization operators is *regularizing* if such a parameter choice exists. A distinction is made between a posteriori parameter choices where $\lambda = \lambda(\delta, g_\delta)$ and a priori parameter choices where $\lambda = \lambda(\delta)$. The latter usually rely on prior information on the problem so that a posteriori choices are preferable in practice.

The above reasoning for Tikhonov regularization extends to a larger class of regularization algorithms. In fact a large class of regularization methods define a one parameter family of regularization operators $R_\lambda : \mathcal{G} \to \mathcal{H}$ to approximate $(A^*A)^{-1}$. For compact operators we can consider

$$f^\lambda_\delta = R_\lambda A^* g = g_\lambda(A^*A) A^* g$$

where the function g_λ defines a regularization operator via

$$R_\lambda = g_\lambda(A^*A) = \sum_{i=1}^{\infty} g_\lambda(\sigma_i^2) \langle \cdot, v_i \rangle_{\mathcal{H}} \, v_i.$$

The same reasoning for the construction of regularization operators for inverse problems defined by compact operators extend to inverse problems defined by bounded operators using spectral theory arguments. Clearly not all the functions g_λ give rise to meaningful algorithms. In regularization theory an abstract definition comprising large class of convergent algorithms is given [EHN96]. The same definition is useful to define regularization algorithms for learning and will be recalled in section 3.3.

It is important for our developments to note that another measure of the error, which is not central in the theory of inverse problems, is the *residual*

$$\left\|Af_\delta^\lambda - Pg\right\|_{\mathcal{G}} = \left\|Af_\delta^\lambda - Af^\dagger\right\|_{\mathcal{G}}. \tag{3.6}$$

The residual will be important in our analysis of learning. We note, comparing (3.5) and (3.6), that the residual is a weaker error measure and clearly while studying the convergence of the residual we do not have to assume the existence of the generalized solution.

Remark 4. *We briefly comment on the well known difference between ill-posedness and ill-conditioning [BDMP88]. Finite dimensional problems are often well-posed. In particular it can be shown that if a solution exists unique then continuity of A^{-1} is always ensured. Nonetheless regularization is needed since the problems are usually ill conditioned and lead to unstable solutions. In fact, if δg and δf are small variations on the data and the solution respectively then we can write*

$$\|\delta f\|_{\mathcal{H}} \leq \left\|A^{-1}\right\| \|\delta g\|_{\mathcal{G}}$$

since A^{-1} is bounded. Moreover since A is bounded we have

$$\|f\|_{\mathcal{H}} \geq \frac{\|g\|_{\mathcal{G}}}{\|A\|}.$$

It follows that the following inequality characterizes the relative variation of the solution w.r.t. the relative variations of the data

$$\frac{\|\delta f\|_{\mathcal{H}}}{\|f\|_{\mathcal{H}}} \leq C(A)\frac{\|\delta g\|_{\mathcal{G}}}{\|g\|_{\mathcal{G}}},$$

where $C(A) = \|A\| \|A^{-1}\|$, namely the conditional number, is always equal or greater then 1. From the above inequality we see that if $C(A) \gg 1$ than we might have unstable solutions even when A^{-1} is continuous.

3.2 Learning as an Inverse Problem

The similarity between regularized least squares and Tikhonov regularization is apparent comparing problems (2.21) and (3.3). However while trying to formalize this analogy several difficulties emerge.

- To treat the problem of learning in the setting of ill-posed inverse problems we have to define a direct problem by means of a suitable operator A between two Hilbert spaces $\mathcal{H}$ and $\mathcal{G}$.

- We have to clarify the relation between consistency, expressed by (2.5), and the convergence considered in (3.5).
- The nature of the noise δ in the context of learning is not clear.

In the following we present a possible way to tackle these problems and show that the problem of learning can be indeed rephrased in a framework close to the one presented in the previous section.

We assume throughout that $\mathcal{H}$ is RKH space with kernel K satisfying assumption 1 and $\ell(y, f(x)) = (y - f(x))^2$. We have already seen that if a hypothesis space $\mathcal{H}$ is given, the *ideal* estimator is the solution of the minimization problem

$$\inf_{f\in\mathcal{H}} \mathcal{E}(f) = \inf_{f\in\mathcal{H}} \|I_K f - f_\rho\|_\rho^2 + \mathcal{E}(f_\rho). \tag{3.7}$$

In the above expression we have stressed the fact that f is an element of $\mathcal{H}$, but its relevant norm is the norm in $L^2(X, \rho_X)$, writing explicitly the embedding operator $I_K : \mathcal{H} \to L^2(X, \rho_X)$. We notice that the action of I_K is trivial since it maps f into itself, but it is not trivial from a topological point of view since it changes the norm from $\|\cdot\|_\mathcal{H}$ to $\|\cdot\|_\rho$. Moreover I_K is continuous since the kernel is bounded (see assumption 1).
Then we simply observe that (3.7) is equivalent to the least square problem associated to the linear inverse problem

$$I_K f = f_\rho. \tag{3.8}$$

We look in $\mathcal{H}$ for a function which is close to the regression function with respect to the ρ-norm.

The fact the above analysis is very natural is apparent if we repeat the above argument replacing the measure ρ with the empirical measure on the sample. In fact in this case we simply recover the standard formulation of a function approximation problem from a fixed grid of input points.
If $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ is a training set then it is straightforward to see that

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 = \min_{f\in\mathcal{H}} \|S_\mathbf{x} f - \mathbf{y}\|_n^2, \tag{3.9}$$

where $\|\cdot\|_n$ is $1/n$ times the euclidean norm in $\mathbb{R}^n$ and $S_\mathbf{x} : \mathcal{H} \to \mathbb{R}^n$ is the sampling operator $(S_\mathbf{x} f)_i = f(x_i)$. Again we can see that empirical risk minimization is the least square problem associated to the linear inverse problem

$$S_\mathbf{x} f = \mathbf{y}. \tag{3.10}$$

As announced we recover the problem of approximating a function from finite data, that is finding f such that $f(x_i) = y_i$ with $i = 1, .., n$.

If we now consider the generalized solutions of the above problem we get some insights into the similarities and differences between the two theories. The Moore-Penrose solution of problem (3.8), if it exists, is nothing but the best in the model for the square loss. To emphasize this fact we denote it with $f_{\mathcal{H}}^{\dagger}$. If $P : L^2(X, \rho_X) \to \mathcal{H}$ is the projection on the closure of $\mathcal{H} \subset L^2(X, \rho_X)$ then $f_{\mathcal{H}}^{\dagger}$ exists if and only if $Pf_\rho \in \mathcal{H}$ and in this case it solves the normal equation $Tf_{\mathcal{H}}^{\dagger} = I_K^* I_K f_{\mathcal{H}}^{\dagger} = I_K^* f_\rho$ and moreover $Pf_\rho = I_K f_{\mathcal{H}}^{\dagger}$. In this case we have that

$$\inf_{f \in \mathcal{H}} \mathcal{E}(f) = \left\| I_K f_{\mathcal{H}}^{\dagger} - f_\rho \right\|_\rho^2 + \mathcal{E}(f_\rho)$$

so that

$$\mathcal{E}(f) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) = \|I_K f - f_\rho\|_\rho^2 - \left\| I_K f_{\mathcal{H}}^{\dagger} - f_\rho \right\|_\rho^2 = \left\| I_K f - I_K f_{\mathcal{H}}^{\dagger} \right\|_\rho^2,$$

where we used the fact that $Pf_\rho = I_K f_{\mathcal{H}}^{\dagger}$, $PI_K f = I_K f$ and $\langle Pf_\rho, f_\rho \rangle_\rho = \langle Pf_\rho, Pf_\rho \rangle_\rho$ If $Pf_\rho \notin \mathcal{H}$ we still have

$$\inf_{f \in \mathcal{H}} \mathcal{E}(f) = \|Pf_\rho - f_\rho\|_\rho^2 + \mathcal{E}(f_\rho)$$

so that reasoning as above we get

$$\mathcal{E}(f) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) = \|I_K f - Pf_\rho\|_\rho^2 .$$

This clarifies two important facts. First, in learning rather than looking at the reconstruction error we are interested into the residual (or better into its square) since this has an interpretation in terms of the expected error. Second, for this reason we no longer need to assume the existence of the generalized solution $f_{\mathcal{H}}^{\dagger}$ (that is the best in the model). This explains the relation between the error considered for consistency in learning and the one considered for convergence in inverse problems. The residual of the solution is a weaker error measure than the reconstruction error usually studied in the inverse problem setting. As we turn to the discrete problem (3.10) we see that the generalized solution in this case is not interesting since it is just a minimal interpolant, i.e. an over-fitting solution. Moreover it is easy to check that the regularized least squares algorithm is simply Tikhonov regularization applied to problem (3.10). Then it is clear that in learning the regularization of the perturbed discrete problem (3.10) should allow us to find a stable solution to the ill-posed continuous problem (3.8). In inverse problems stability to the noise and the effects of discretization are usually treated separately. In the learning framework this cannot be done because of the random sampling and regularization should take care of both aspects. To make this last fact precise it is still left to discuss the different type of convergence analysis: with respect to the noise in inverse problems and with respect to the number of data in learning. A better understanding of this relation can be achieved trying to define a measure of the perturbation relating problem (3.8) and (3.10).

3.2.1 Stochastic Perturbation Measures

As we try to compare problems (3.8) and (3.10) we notice that the operators defining the two problems have range in different spaces. This is because in the learning setting the discretization procedure is not controlled since it is due to random sampling.

A possible way to find a measure of the perturbation due to the sampling relies on noting that that the least square solutions of (3.7) and (3.9) are solutions of the following linear equations

$$I_K^* I_K f = I_K^* f_\rho, \qquad S_{\mathbf{x}}^* S_{\mathbf{x}} f = S_{\mathbf{x}}^* \mathbf{y}. \tag{3.11}$$

In the above formulation $I_K^* I_K$ and $S_{\mathbf{x}}^* S_{\mathbf{x}}$ are operators from $\mathcal{H}$ to $\mathcal{H}$, whereas $I_K^* f_\rho$ and $S_{\mathbf{x}}^* \mathbf{y}$ are elements of $\mathcal{H}$. This suggests that the perturbation measure due to random sampling can be expressed by the quantities $\|I_K^* f_\rho - S_{\mathbf{x}}^* \mathbf{y}\|_{\mathcal{H}}$ and $\|I_K^* I_K - S_{\mathbf{x}}^* S_{\mathbf{x}}\| = \|T - T_{\mathbf{x}}\|$ which are clearly random variables. Intuitively if we look at the explicit form of the quantities appearing in the above equations we might expect $S_{\mathbf{x}}^* S_{\mathbf{x}}$ and $S_{\mathbf{x}} \mathbf{y}$ to converge to $I_K^* I_K$ and $I_K^* f \rho$ respectively, when the number n of data goes to infinity, as a consequence of the law of large numbers. This is formalized by the following propositions.

Proposition 2. *Let K satisfy assumption 1.*

- *Then for all $n \in \mathbb{N}$ and $0 < \eta < 1$,*

$$\Pr\left(\|T - T_{\mathbf{x}}\| \leq \frac{1}{\sqrt{n}} 2\sqrt{2}\kappa^2 \sqrt{\log \frac{2}{\eta}} \right) \geq 1 - \eta. \tag{3.12}$$

- *If $Y = [-M, M]$, for some $M > 0$, then for all $n \in \mathbb{N}$ and $0 < \eta < 1$*

$$\Pr\left(\|I_K^* f_\rho - S_{\mathbf{x}}^* \mathbf{y}\| \leq \frac{1}{\sqrt{n}} 2\sqrt{2}\kappa M \sqrt{\log \frac{2}{\eta}} \right) \geq 1 - \eta. \tag{3.13}$$

- *If $f_{\mathcal{H}}^\dagger$ exists and for some $\Sigma, M \in \mathbb{R}^+$ it holds*

$$\int_Y \left(e^{\frac{|y - f_{\mathcal{H}}^\dagger(x)|}{M}} - \frac{|y - f_{\mathcal{H}}^\dagger(x)|}{M} - 1 \right) d\rho(y|x) \leq \frac{\Sigma^2}{2M^2}, \qquad \text{for } \rho_X\text{-almost all } x \in X \tag{3.14}$$

then for all $n \in \mathbb{N}$ and $0 < \eta < 1$

$$\Pr\left(\left\| T_{\mathbf{x}} f_{\mathcal{H}}^\dagger - S_{\mathbf{x}}^* \mathbf{y} \right\|_{\mathcal{H}} \leq 2\left(\frac{\kappa M}{n} + \frac{\kappa \Sigma}{\sqrt{n}}\right) \log \frac{2}{\eta} \right) \geq 1 - \eta. \tag{3.15}$$

The above concentration inequalities give the desired perturbation measures. Such perturbation measures go to zero as the number of examples goes to infinity.

The proof of the above proposition relies on the following concentration results for Hilbert space valued random variables [PS85].

Lemma 1. *Let $(\Omega, \mathcal{B}, P)$ be a probability space and ξ a random variable on Ω with values in a real separable Hilbert space $\mathcal{G}$. Assume there are two constants H, σ such that*

$$\mathrm{E}\left[\|\xi - \mathrm{E}[\xi]\|_{\mathcal{G}}^m\right] \leq \frac{1}{2} m! \sigma^2 H^{m-2}, \quad \forall m \geq 2 \tag{3.16}$$

then, for all $n \in \mathbb{N}$ and $0 < \eta \leq 1$,

$$\Pr\left(\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i - \mathrm{E}[\xi] \right\|_{\mathcal{G}} \leq 2\left(\frac{H}{n} + \frac{\sigma}{\sqrt{n}}\right) \log \frac{2}{\eta} \right) \geq 1 - \eta.$$

In particular (3.16) is verified if we have $\mathrm{E}[\xi] \leq H$ and $\mathrm{E}[\|\xi\|_{\mathcal{H}}^2] \leq \sigma^2$.

For bounded random variables we can use the following simplified results.

Corollary 2. *Let $(\Omega, \mathcal{B}, P)$ be a probability space and (ξ) a zero mean random variable on Ω with values in a real separable Hilbert space $\mathcal{G}$. If $\|\xi\|_{\mathcal{G}} \leq C \leq \infty$, then for all $n \in \mathbb{N}$ and $0 < \eta \leq 1$,*

$$\Pr\left(\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\|_{\mathcal{G}} \leq \frac{\sqrt{2}C}{\sqrt{n}} \sqrt{\log \frac{2}{\eta}} \right) \geq 1 - \eta$$

We are now ready to prove proposition 2.

Proof of proposition 2. The first two items can be easily proved using corollary 2. For inequality (3.12) it is crucial to remember that $T, T_{\mathbf{x}} \in \mathcal{B}_2(\mathcal{H})$ the space of Hilbert Schmidt operators and that $\|\cdot\| \leq \|\cdot\|_2$. Then we can define the random variable $\xi : X \to \mathcal{B}_2(\mathcal{H})$ such that $\xi = \langle \cdot, K_x \rangle_{\mathcal{H}} K_x - T$ and it is straightforward to check that

$$\mathrm{E}[\xi] = 0, \qquad \|\xi\|_2 \leq 2\kappa^2$$

so that the proof follows applying corollary 2. Similarly for inequality (3.13) we can define $\xi : Z \to \mathcal{H}$ such that $\xi = yK_x - I_K^* f_\rho$ and

$$\mathrm{E}[\xi] = 0, \qquad \|\xi\|_{\mathcal{H}} \leq 2\kappa M$$

so that again applying corollary 2 we get the proof. Finally to prove inequality (3.15) we consider the random variable $\xi : Z \to \mathcal{H}$ defined by

$$\xi = K_x(y - f_{\mathcal{H}}^{\dagger}(x)).$$

It easy to prove that ξ is a zero mean random variable, in fact

$$\begin{aligned}
\mathrm{E}\left[\xi\right] &= \int_{X\times Y} K_x y - K_x \left\langle f_{\mathcal{H}}^{\dagger}, K_x \right\rangle_{\mathcal{H}} d\rho(x,y) \\
&= \int_X d\rho_X(x) K_x \left(\int_Y y d\rho(y|x)\right) - \int_X \left\langle f_{\mathcal{H}}^{\dagger}, K_x \right\rangle_{\mathcal{H}} K_x d\rho_X(x) \\
&= I_K^* f_\rho - T f_{\mathcal{H}}^{\dagger}.
\end{aligned}$$

Recalling that $f_{\mathcal{H}}^{\dagger}$ solves the normal equation (3.11) we have that $T f_{\mathcal{H}}^{\dagger} = I_K^* f_\rho$ so that the above mean is zero. Moreover assumption (3.14) ensures (see for example [vdVW96])

$$\int_Y (y - f_{\mathcal{H}}^{\dagger}(x))^m d\rho(y|x)) \leq \frac{1}{2} m! \Sigma^2 M^{m-2}, \quad \forall m \geq 2$$

so that

$$\begin{aligned}
\mathrm{E}\left[\|\xi\|_{\mathcal{H}}^m\right] &= \int_{X\times Y} \left(\left\langle K_x(y - f_{\mathcal{H}}^{\dagger}(x)), K_x(y - f_{\mathcal{H}}^{\dagger}(x)) \right\rangle_{\mathcal{H}} \right)^{\frac{m}{2}} d\rho(x,y) \\
&= \int_X d\rho_X(x) K(x,x) \int_Y (y - f_{\mathcal{H}}^{\dagger}(x))^2 d\rho(y|x)) \\
&\leq \kappa^m \frac{1}{2} m! \Sigma^2 M^{m-2} \leq \frac{1}{2} m! (\kappa\Sigma)^2 (\kappa M)^{m-2}.
\end{aligned}$$

The proof follows applying (1) with $H = \kappa M$ and $\sigma = \kappa\Sigma$. □

3.3 Abstract Characterization of Regularization

In this section we describe how to derive and define regularization for learning. The content of this section is divided into two parts. First we discuss how the construction of regularization operators extend to the setting of learning. Second we give an abstract characterization of regularization by means of a set of simple conditions.

3.3.1 Regularization Operators for Learning

Let us recall that in the formulation of the learning problem discussed in section 3.2 we have that problems (3.8) and (3.10) induce the two normal equations (3.11). Our goal is to solve an ill-posed continuous problem given a discretization of it. Since we do not control the discretization we demand the regularization algorithms to take care of such an indetermination.

In our setting $T = I_K^* I_K$ is a compact operator so that in general the (Moore-Penrose) inverse of I_K is not continuous and hence finding the generalized solution $f_{\mathcal{H}}^{\dagger}$ is an ill-posed problem. Moreover we only have the perturbed $T_{\mathbf{x}} = S_{\mathbf{x}}^* S_{\mathbf{x}}$ in place of T (and $S_{\mathbf{x}}^* \mathbf{y}$ in place of $I_K^* f_\rho$).

As we discussed in section 3.1 the key idea of inverse problems is to regularize (3.8) by considering a family of regularized solutions

$$f^\lambda = g_\lambda(I_K^* I_K) I_K^* f_\rho \tag{3.17}$$

depending on a positive parameter λ in such a way that $g_\lambda(I_K^* I_K)$ is a family of operators approximating the inverse of I_K when λ goes to 0. When we work on the data (3.17) is replaced by

$$f_{\mathbf{z}}^\lambda = g_\lambda(S_{\mathbf{x}}^* S_{\mathbf{x}}) S_{\mathbf{x}}^* \mathbf{y}. \tag{3.18}$$

The idea is then that if we choose λ correctly we can ensure that (3.18) is close to (3.17) and (3.17) is a good proxy for $f_{\mathcal{H}}^{\dagger}$. Indeed we prove the above intuition in chapter 4. The final estimator is then defined providing the above scheme with a suitable parameter choice $\lambda_n = \lambda(n, \mathbf{z})$ so that $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$.

From the algorithmic point of view it is easy to see that the above regularization always induce suitable kernel methods. Each regularization scheme defines a corresponding kernel method by means of

$$f_{\mathbf{z}}^\lambda(x) = \sum_{i=1}^n \alpha_i K(x, x_i) \qquad \text{with} \qquad \alpha = \frac{1}{n} g_\lambda(\frac{\mathbf{K}}{n})\mathbf{y} \tag{3.19}$$

and again the final estimator is defined providing the above scheme with a parameter choice $\lambda_n = \lambda(n, \mathbf{z})$ so that $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$.

Indeed, we can prove (3.19) recalling that by polar decomposition the following equalities hold $S_{\mathbf{x}} = \sqrt{1/n\mathbf{K}} U_{\mathbf{x}}^*$, $S_{\mathbf{x}}^* = U_{\mathbf{x}} \sqrt{1/n\mathbf{K}}$ and clearly $T_{\mathbf{x}} = U_{\mathbf{x}} 1/n\mathbf{K} U_{\mathbf{x}}^*$. Then we can write

$$f_{\mathbf{z}}^\lambda = g_\lambda(T_{\mathbf{x}}) S_{\mathbf{x}}^* y = U_{\mathbf{x}} g_\lambda(\frac{1}{n}\mathbf{K}) \sqrt{\frac{1}{n}\mathbf{K}} y \tag{3.20}$$

where we used the fact that $U_{\mathbf{x}}^* U_{\mathbf{x}}$ is the identity on the range of $\mathbf{K}$. From the above formula we immediately see that $f_{\mathbf{z}}^\lambda$ is an element of the range of $U_{\mathbf{x}}$, which is the linear span of the vectors K_{x_i}. Hence $f_{\mathbf{z}}^\lambda = \sum_{i=1}^n \alpha_i K_{x_i}$ and, if we apply the sampling operator on both sides of (3.20), we get

$$S_{\mathbf{x}} f_{\mathbf{z}}^\lambda = S_{\mathbf{x}} \sum_{i=1}^n \alpha_i K_{x_i} = \mathbf{K}\alpha$$

where α denotes the vector of the coefficients and

$$S_{\mathbf{x}}U_{\mathbf{x}}g_\lambda(\frac{1}{n}\mathbf{K})\sqrt{\frac{1}{n}\mathbf{K}}y = \sqrt{\frac{1}{n}\mathbf{K}}g_\lambda(\frac{1}{n}\mathbf{K})\sqrt{\frac{1}{n}\mathbf{K}}y.$$

Then the following equality holds

$$\mathbf{K}\alpha = \frac{1}{n}\mathbf{K}g_\lambda(\frac{1}{n}\mathbf{K})y$$

and (3.19) easily follows.

Clearly not all the functions g_λ are admissible and we give a characterization of regularization in the next section.

3.3.2 Definition of Regularization

We now present the class of regularization algorithms we are going to study. Regularization is defined according to what is usually done for ill-posed inverse problems. We will show that the following definition characterizes which regularization provide sensible learning algorithms.

Definition 5 (Regularization). *We say that a family $g_\lambda : [0, \kappa^2] \to \mathbb{R}$, $0 < \lambda \leq \kappa^2$, is regularization if the following conditions hold*

- *There exists a constant D such that*

$$\sup_{0<\sigma\leq\kappa^2} |\sigma g_\lambda(\sigma)| \leq D \tag{3.21}$$

- *There exists a constant B such that*

$$\sup_{0<\sigma\leq\kappa^2} |g_\lambda(\sigma)| \leq \frac{B}{\lambda} \tag{3.22}$$

- *The qualification of the regularization g_λ is the maximal ν such that*

$$\sup_{0<\sigma\leq\kappa^2} |1 - g_\lambda(\sigma)\sigma|\sigma^\nu \leq \gamma_\nu \lambda^\nu, \tag{3.23}$$

  where γ_ν does not depend on λ.

Let us briefly discuss such conditions. The first two conditions basically ensure that the obtained algorithm can be seen as family of linear continuous maps, parameterized by the regularization parameter λ. In fact, since $g_\lambda(\sigma)$ is bounded for σ in $[0, \kappa^2]$, then the spectral theorem ensures that $g_\lambda(I_K^* I_K)$ is bounded, too.
The second condition also ensures that the solution of the population problem converges to the best in the model when λ goes to zero (or that $I_K f^\lambda$ converges to Pf_ρ). In other words this ensures that the bias (approximation error) goes to zero as λ goes to zero. In fact $g_\lambda(\sigma)$ approximates the function $\frac{1}{\sigma}$ as λ goes to 0, that is, $g_\lambda(I_K^* I_K)$ is a family of operators approximating the inverse of $I_K^* I_K$ when λ goes to 0.
The third condition allows to derive the convergence rate for the approximation error if $f_{\mathcal{H}}^\dagger$ (or Pf_ρ) satisfies suitable a priori conditions. The notion of qualification quantifies the capability of a given algorithm of exploiting the regularity of the target function. The performance of methods with finite qualification no longer improves beyond a certain regularity level. In the following we discuss in some more detail the relation between qualification and a priori conditions. Finally we note that the various constant appear in the bounds and will be different for each algorithm.

3.4 Regularization Algorithms

We now list several algorithms that fall into above definition. From (3.19) we know that each algorithm is a kernel method whose solution can be represented in closed form.

3.4.1 Tikhonov Regularization

We start our discussion giving a brief review of Tikhonov regularization. In this case the regularization is $g_\lambda(\sigma) = \frac{1}{\sigma+\lambda}$. It is straightforward to check that (3.21) and (3.22) hold with $B = D = 1$. Condition (3.23) is verified with $\gamma_\nu = 1$ for $0 < \nu \leq 1$ and hence the qualification is equal to 1.
According to (3.19) the algorithm amounts to a matrix inversion problem since we have to solve

$$\alpha = (\mathbf{K} + n\lambda I)^{-1}\mathbf{y}.$$

Such an algorithm is well known and can be written as the variational problem

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \} \tag{3.24}$$

which is simply the regularization network [EPP00] induced by the square loss. For the linear kernel the above algorithm is known as ridge regression in statistics.

3.4.2 Landweber Iteration and Gradient Descent Learning

Landweber iteration is characterized by

$$g_t(\sigma) = \tau \sum_{i=0}^{t-1} (1 - \tau\sigma)^i$$

where we identify $\lambda = t^{-1}, t \in \mathbb{N}$ and take $\tau = 1/\kappa^2$. In this case we have $B = D = 1$ and the qualification is infinite since (3.23) holds with $\gamma_\nu = 1$ if $0 < \nu \leq 1$ and $\gamma_\nu = \nu^\nu$ otherwise (note that γ_ν grows larger with ν). From the algorithmic point of view we can rewrite the algorithm as the following iterative map

$$\alpha_i = \alpha_{i-1} + \frac{\tau}{n}(\mathbf{y} - \mathbf{K}\alpha_{i-1}), \qquad i = 1, \ldots, t-1 \tag{3.25}$$

setting $\alpha_0 = 0$. We now give some comments and remark. First we recall that in [YRC05] a more general form of the relaxation parameter is considered, namely the variable step-size

$$\tau_i = \frac{1}{\kappa^2 (i+1)^\theta} \tag{3.26}$$

with $0 \leq \theta < 1$. Interestingly it was shown that the fixed step-size $\tau = 1/\kappa^2$ is the best choice among the variable step-size in (3.26). This suggests that τ does not play any role for regularization. We review this results in next chapter.

Second we note that the above algorithm has an easy interpretation allowing connections with other algorithms.
In fact we can see that the above instance of Landweber iteration simply corresponds to the minimization of the empirical risk via gradient descent. In fact recalling that the Gatéaux derivative of a functional $\Phi : \mathcal{H} \to \mathbb{R}$ at a point f is defined as the map $D_f : \mathcal{H} \to \mathbb{R}$ such that

$$D_f\Phi(h) = \frac{d\Phi(f + th)}{dt}\Big|_{t=0}, \qquad h \in \mathcal{H}$$

we have that the derivative of the empirical risk is

$$D\mathcal{E}_{\mathbf{z}}(f)(h) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - f(x_i))h(x_i) = \left\langle -\frac{2}{n}\sum_{i=1}^{n}(y_i - f(x_i))K_{x_i}, h \right\rangle_{\mathcal{H}}$$

so that we find the explicit form of the gradient empirical risk as

$$\nabla\mathcal{E}_{\mathbf{z}}(f) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - f(x_i))K_{x_i}. \tag{3.27}$$

If we write explicitly $f_{\mathbf{z}}^t = \tau \sum_{i=0}^{t-1} (I - \tau S_{\mathbf{x}}^* S_{\mathbf{x}})^i S_{\mathbf{x}}^* \mathbf{y}$ as an iterative map we get

$$f_{\mathbf{z}}^{i+1} = f_{\mathbf{z}}^i + \frac{\tau}{n} \sum_{j=1}^{n} (y_j - f_{\mathbf{z}}^i(x_j)) K_{x_j}, \qquad f_{\mathbf{z}}^0 = 0, \tag{3.28}$$

and looking at (3.27) we see that (3.28) is exactly the gradient descent minimization of $\mathcal{E}_{\mathbf{z}}(f)$ in the RKH space $\mathcal{H}$.
In the perspective of a greedy minimization of the empirical error the above algorithm has some interesting connection to boosting algorithms.

3.4.2.1 Connection to Boosting

The notion of *boosting* was originally proposed in connection to the question of weather a "weak" learning algorithm which performs just slightly better than random guessing (with success probability slightly larger than 1/2) can be "boosted" into a "strong" learning algorithm of high accuracy [Val84]. The AdaBoost algorithm [FS97] is probably the more prominent example in this class of algorithms since it proved to be extremely effective in practice. Recently it was shown that many boosting algorithms can be seen as a greedy procedure to minimize the empirical error [MBBF00]. Many different algorithms can be recovered in a general framework changing the loss function and the hypotheses space. In order to understand such algorithms it is crucial to devise the regularization principle at the basis of them. Indeed it turns out that the greedy minimization of the empirical error is often subject to some additional constraint [ZY03] which is explicitly imposed to the complexity of the solution. Considering the square loss and hypotheses spaces which are RKH spaces, we show that the early stopping of the iteration procedure has a self-regularizing effect and can be seen (a posteriori) to restrict the solution into a ball in the RKH space. *The minimization is done without any constraints but the early stopping of the iteration.* In particular our algorithm can be seen as a generalization of the L2-Boost algorithm [BY02] where splines were considered as hypotheses spaces and the connection to Landweber iteration was not observed. In the light of boosting we can consider the functions K_x as our weak learners.

3.4.3 Semiiterative Regularization and the ν-method

An interesting class of algorithms are the so called semiiterative regularization or accelerated Landweber iteration. This class of methods can be seen as a generalization of Landweber iteration where the regularization is now

$$g_t(\sigma) = p_t(\sigma)$$

with p_t polynomial of degree $t-1$. In this case we can identify $\lambda = t^{-2}, t \in \mathbb{N}$ and we assume $\kappa = 1$ for simplicity. We have $B = 2$ and $D = 1$. The qualification of this class of method is usually finite. An example which turns out to be particularly interesting is the so called $\nu - method$. We refer to [EHN96] for a derivation of this method. In the $\nu - method$ the qualification is ν (fixed) with $\gamma_\nu = c$ for some positive constant c. The algorithm amounts to solving, for $\alpha_0 = 0$, the following map

$$\alpha_i = \alpha_{i-1} + u_i(\alpha_{i-1} - \alpha_{i-2}) + \frac{\omega_i}{n}(\mathbf{y} - \mathbf{K}\alpha_{i-1}), \qquad i = 1, \ldots, t-1$$

where

$$\begin{aligned} u_i &= \frac{(i-1)(2i-3)(2i+2\nu-1)}{(i+2\nu-1)(2i+4\nu-1)(2i+2\nu-3)} \\ \omega_i &= 4\frac{(2i+2\nu-1)(i+\nu-1)}{(i+2\nu-1)(2i+4\nu-1)} \qquad t > 1. \end{aligned}$$

The interest of this method lies in the fact that since the regularization parameter here is $\lambda = t^{-2}$ we just need the square root of the number of iterations needed by Landweber iteration. In inverse problems this method proved to be extremely fast and is often used as valid alternative to conjugate gradient (see [EHN96], Chapter 6 for details).

3.4.4 Spectral Cut-off

A classical regularization algorithms for ill-posed inverse problems is spectral cut-off or truncated singular value decomposition (TSVD) defined by

$$g_\lambda = \begin{cases} \frac{1}{\sigma}, & \sigma \geq \lambda \\ 0, & \sigma < \lambda \end{cases} .$$

This method up-to our knowledge is not used in Learning Theory. In this case we have $B = D = 1$. The qualification of the method is arbitrary and $\gamma_\nu = 1$. From the algorithmic point of view it amounts to consider the truncated singular values decomposition of the kernel matrix.

3.4.5 Iterated Tikhonov

As we have seen while discussing Tikhonov regularization such method has finite qualification and this reflects in the impossibility to exploit the regularity of the solution beyond

a certain regularity level. To overcome this problem the following regularization can be considered

$$g_{\lambda,t}(\sigma) = \frac{(\sigma+\lambda)^t - \sigma^t}{\lambda(\sigma+\lambda)^t}.$$

In this case $D = 1$ and $B = t$ and the qualification of the method is now t with $\gamma_\nu = 1$. The algorithm is described by the following iterative map

$$(\mathbf{K} + n\lambda I)\alpha_i = \mathbf{y} + n\lambda\alpha_{i-1} \qquad i = 1, \ldots, t$$

choosing $\alpha_0 = 0$. It is easy to see that for $t = 1$ we simply recover the standard Tikhonov regularization but as we let $t > 0$ we improve the qualification of the method. Moreover we note that by fixing λ we can think of the above algorithms as an iterative regularization with t regularization parameter.

3.5 Filter Function Perspective

Similarly to what is done in inverse problems we can give an interpretation of regularization from a filter function point of view. We start our discussion focusing on Tikhonov regularization or regularized least-squares algorithm (RLSA) defining a family of estimators via regularized least square problem (3.24) depending on the regularization parameter λ. The final estimator is defined providing the above scheme with a parameter choice $\lambda_n = \lambda(n, \mathbf{z})$ so that $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$. Understanding the way such an algorithm works allows to derive different regularization schemes. As we have seen in Chapter 2 a possible interpretation of RLSA relates the penalty $\|f\|_{\mathcal{H}}^2$ to the complexity of the solution. that is we look at the RLSA as an approximate implementation of Structural Risk Minimization.

Another point consists in considering the penalty term as a smoothness term which enforces stability of the solution where stability is again with respect to the random sampling of the data. Although this point of view is not new to learning theory since the connection between stability and generalization was considered in [BE02, MNPR04, PRMN04], as we restrict our analysis to the quadratic loss function we can have some interesting insight.

Indeed the regularized least-squares algorithm can be seen as implementing a low pass filter on the expansion of the regression function on suitable basis. We have already seen that the solution of problem (3.24) can be written as

$$f_{\mathbf{z}}^{\lambda}(x) = \sum_{i=1}^{n} \alpha K(x, x_i), \qquad \alpha = (\mathbf{K} + n\lambda I)^{-1}\mathbf{y}, \tag{3.29}$$

where $\mathbf{K}$ is the kernel matrix $(\mathbf{K})_{ij} = K(x_i, x_j)$. Such result is known as representer theorem [KW70] and we implicitly give a proof of it in section 3.3.1.

From the explicit form of the coefficients we see that as $\lambda > 0$ we are numerically stabilizing a matrix inversion problem which is possibly ill-conditioned (that is numerically unstable). This is important from the algorithmic point of view, but it is also crucial to ensure the generalization properties of the estimator. When ρ is known we have the population version of (3.24)

$$\min_{f\in\mathcal{H}}\{\int_{X\times Y}(y-f(x))^2 d\rho(x,y)+\lambda\left\|f\right\|_{\mathcal{H}}^2\}. \tag{3.30}$$

A generalized representer theorem[3] (see for example [CS02a]) gives the explicit form of the solution as

$$f^\lambda = (L_K+\lambda I)^{-1}L_K f_\rho$$

where L_K is the integral operator of kernel K acting in $L^2(X,\rho_X)$ (see section 2.3.2) and we considered f^λ as a function in $L^2(X,\rho_X)$. We noted in chapter 2 that since the kernel is bounded, symmetric and positive definite, L_K is a positive compact operator and the spectral theorem ensures the existence of a basis of eigenfunctions $L_K u_i = \sigma_i^2 u_i$ with $\sigma_i^2 \geq 0$. Then we can rewrite the solution of the above problem as

$$f^\lambda = \sum_{i=1}^{\infty}\frac{\sigma_i^2}{\sigma_i^2+\lambda}\left\langle f_\rho, u_i\right\rangle_\rho u_i.$$

From the latter expression we see that the effect of regularization is that of a low pass filter which selects the components of the regression function corresponding to large eigenvalues. If we slightly perturb ρ, the operator L_K and f_ρ change, however the filter ensures that corresponding solution f^λ is close to f_ρ provided that the perturbation is small and the parameter λ is suitably chosen. *The idea is that we can look to the sample case exactly as a perturbation on the problem due to random sampling.* In this case we think of $\mathbf{y}$ and $\mathbf{K}$ as perturbations of f_ρ and L_K respectively. The low pass filter is then a way to ensure stability and it is natural to extend this approach to other *regularization* g_λ such that

$$f^\lambda = \sum_{i=1}^{\infty} g_\lambda\sigma_i^2\left\langle f_\rho, u_i\right\rangle_\rho u_i.$$

3.6 A Priori Assumption and General Source Condition

As we already mentioned to obtain probabilistic bounds such as that in (2.4) (or rather bounds on (2.7)) we have to restrict the class of possible probability measures. In Learning Theory this is related to the so called "no free lunch" theorem [DGL96] but similar kind

[3]Again we implicitly gave a proof of this fact in section 3.3.1.

of phenomenon occurs in statistics [GKW96] and in regularization of ill-posed inverse problems [EHN96]. Essentially what happens is that we can always find a solution with convergence guarantees to some prescribed target function but the convergence rates can be arbitrary slow. In our setting this turns into the impossibility to state finite sample bounds holding uniformly with respect to any probability measure ρ.

A standard way to impose restrictions on the class of possible problems is to consider a set of probability measures $\mathcal{M}(\Omega)$ such that the associated regression functions satisfy $f_\rho \in \Omega$. Such a condition is called the prior. The set Ω is usually a compact subset of $L^2(X, \rho_X)$ determined by smoothness conditions [DKPT04]. In the context of RKH spaces it is natural to describe the prior in term of the compact operator L_K, considering $f_\rho \in \Omega_{r,R}$ with

$$\Omega_{r,R} = \{f \in L^2(X, \rho_X) : f = L_K^r w, \|w\|_\rho \leq R\}. \tag{3.31}$$

The above condition is often written as $\left\|L_K^{-r} f_\rho\right\|_\rho \leq R$ [CS02b, SZ05]. Note that, when $r = 1/2$, such a condition is equivalent to assume $f_\rho \in \mathcal{H}$ and is independent of the measure ρ, but for arbitrary r it is distribution dependent.

As noted in [DVRC$^+$05b, DVRC05a] the condition $f_\rho \in \Omega_{r,R}$ corresponds to what is called a source condition in the inverse problems literature. In fact if we consider $Pf_\rho \in \Omega_{r,R}$, $r > 1/2$, then $Pf_\rho \in Im(I_K)$ and we can equivalently consider the prior $f_\mathcal{H}^\dagger \in \Omega_{c,R}$ with

$$\Omega_{c,R} = \{f \in \mathcal{H} : f = T^c h, \|h\|_\mathcal{H} \leq R\} \tag{3.32}$$

where $c = r - 1/2$. In fact if we let $I_K^* = U(I_K I_K^*)^{\frac{1}{2}}$ be the polar decomposition of I_K^*, then for $r > 1/2$

$$f_\rho = (I_K I_K^*)^r \phi = (I_K I_K^*)^{1/2} (I_K I_K^*)^c \phi = (I_K I_K^*)^{1/2} U^* U (I_K I_K^*)^c U^* U \phi = I_K (T)^c U \phi,$$

where $c = r - 1/2$. It follows that $Pf_\rho \in Im(I_K)$, so that $f_\mathcal{H}^\dagger$ exists and since $Pf_\rho = I_K f_\mathcal{H}^\dagger$ clearly $f_\mathcal{H} = (T)^c U \phi$.

Recalling that $T = I_K^* I_K$ we see that the above condition is the standard source condition for the linear problem $I_K f = f_\rho$, namely Hölder source condition [EHN96].

Following what is done in inverse problems in this paper we wish to extend the class of possible probability measures $\mathcal{M}(\Omega)$ considering general source condition (see [MP03] and references therein). We illustrate this approach under the assumption $Pf_\rho \in R(I_K)$, that is when the best in model exists $f_\mathcal{H}^\dagger$ exists but similar idea can be used to extend (3.31) in the case when $f_\mathcal{H}^\dagger$ does not exist. If $f_\mathcal{H}^\dagger$ exists then it solves the normalized embedding equation $Tf = I_K^* f_\rho$. Using the singular value decompositions

$$T = \sum_{i=1}^\infty \sigma_i^2 \langle \cdot, v_i \rangle_\mathcal{H} v_i \qquad L_K = \sum_{i=1}^\infty \sigma_i^2 \langle \cdot, u_i \rangle_\rho u_i,$$

for orthonormal systems $\{v_i\}$ in $\mathcal{H}$ and $\{u_i\}$ in $L^2(X, \rho_X)$ and a sequence of singular numbers $\kappa^2 \geq \sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq 0$, one can represent $f_{\mathcal{H}}^{\dagger}$ in the form

$$f_{\mathcal{H}}^{\dagger} = \sum_{i=1}^{\infty} \frac{1}{\sigma} \langle f_\rho, u_i \rangle_\rho \, v_i.$$

Then $f_{\mathcal{H}}^{\dagger} \in \mathcal{H}$ if and only if

$$\sum_{i=1}^{\infty} \frac{\langle f_\rho, u_i \rangle_\rho^2}{\sigma_i^2} < \infty.$$

The above condition is known as Picard's criterion and provides a zero-smoothness condition on $f_{\mathcal{H}}^{\dagger}$ (merely $f_{\mathcal{H}}^{\dagger} \in \mathcal{H}$) which tells us that the Fourier coefficients $\langle f_\rho, u_i \rangle_\rho$ should decay much faster than σ_i^2. Therefore it seems natural to measure the smoothness of $f_{\mathcal{H}}^{\dagger}$ by enforcing some faster decay. More precisely, not only Picard's criterion but also the stronger condition

$$\sum_{i=1}^{\infty} \frac{\langle f_\rho, u_i \rangle_\rho^2}{\sigma_i^2 \phi^2(\sigma_i^2)} < \infty$$

is satisfied, where ϕ is some continuous increasing function defined on the interval $[0, \kappa^2] \supset \{\sigma_i^2\}$ and such that $\phi(0) = 0$. Then

$$h := \sum_{i=1}^{\infty} \frac{1}{\sqrt{\sigma_i^2}\phi(\sigma_i^2)} \langle f_\rho, u_i \rangle_\rho \, v_i$$

and

$$f_{\mathcal{H}}^{\dagger} = \sum_{i=1}^{\infty} \phi(t_i) \langle h, v_i \rangle \, v_i = \phi(T)h \in \mathcal{H}.$$

Thus, additional smoothness of $f_{\mathcal{H}}^{\dagger}$ can be expressed as an inclusion

$$f_{\mathcal{H}}^{\dagger} \in \Omega_{\phi,R} := \{f \in \mathcal{H} : f = \phi(T)h, \|h\|_{\mathcal{H}} \leq R\}, \tag{3.33}$$

that goes usually under the name of *source condition*. The function ϕ is called index function. As we discuss, in more details, in the following we always consider index function admitting a decomposition

$$\phi(\sigma) = \theta(\sigma)\psi(\sigma) \tag{3.34}$$

where is ψ is operator concave and θ Lipschitz. The source conditions we described at the beginning of this section and many others are modeled by the above class of index functions (see remark 5). Moreover there is a good reason to restrict the class of possible index functions as we did in (3.34). In fact, in general the smoothness expressed through source conditions is not stable with respect to perturbations in the involved operator T. In

Learning Theory only the empirical covariance operator $T_{\mathbf{x}}$ is available and it is desirable to control $\phi(T) - \phi(T_{\mathbf{x}})$. This can be achieved by requiring ϕ to be operator monotone. Recall that the function ϕ is operator monotone on $[0, b]$ if for any pair of self-adjoint operators U, V, with spectra in $[0, b]$ such that $U \leq V$ we have $\phi(U) \leq \phi(V)$. The partial ordering $B_1 \leq B_2$ for self-adjoint operators B_1, B_2 on some Hilbert space $\mathcal{H}$ means that for any $h \in \mathcal{H}$, $\langle B_1 h, h\rangle \leq \langle B_2 h, h\rangle$. It follows from the Löwner theorem (see for example [Han00]) that each operator monotone function on $(0, b)$ admits an analytic continuation in the corresponding strip of the upper half-plane with positive imaginary part. Important implications of the concept of operator monotonicity in the context of regularization can be seen from the following result (see [MP02, MP05a]).

Theorem 3. *Suppose ψ is an operator monotone index function on $[0, b]$, with $b > a$. Then there is a constant $c_\psi < \infty$ depending on $b - a$, such that for any pair B_1, B_2, $\|B_1\|, \|B_2\| \leq a$, of non-negative self-adjoint operators on some Hilbert space it holds*

$$\|\psi(B_1) - \psi(B_2)\| \leq c_\psi \psi(\|B_1 - B_2\|).$$

Moreover, there is $c > 0$ such that

$$c\frac{\lambda}{\psi(\lambda)} \leq \frac{\sigma}{\psi(\sigma)}$$

whenever $0 < \lambda < \sigma \leq a < b$.

Thus operator monotone index functions allow a desired norm estimate for $\phi(T) - \phi(T_{\mathbf{x}})$. Therefore in the following we consider index functions from the class

$$\mathcal{F}_C = \{\psi : [0, b] \to \mathbb{R}_+, \text{operator monotone}, \psi(0) = 0, \psi(b) \leq C, b > \kappa^2\} \tag{3.35}$$

Note that from the above theorem it follows that an index function $\psi \in \mathcal{F}_C$ cannot converge faster than linearly to 0. To overcome this limitation of the class $\mathcal{F}_C$ we also introduce the class $\mathcal{F}$ of index functions $\phi : [0, \kappa^2] \to \mathbb{R}_+$ which can be split into a part $\psi \in \mathcal{F}_C$ and a monotone Lipschitz part $\vartheta : [0, \kappa^2] \to \mathbb{R}_+$, $\vartheta(0) = 0$, i.e. $\phi(\sigma) = \vartheta(\sigma)\psi(\sigma)$. This splitting is not unique such that we implicitly assume that the Lipschitz constant for ϑ is equal to 1 which means

$$\|\vartheta(T) - \vartheta(T_{\mathbf{x}})\| \leq \|T - T_{\mathbf{x}}\|.$$

The fact that an operator valued function ϑ is Lipschitz continuous if a real function ϑ is Lipschitz continuous follows from theorem 8.1 in [BS03].

Remark 5. *Observe that for $c \in [0, 1]$ a Hölder-type source condition (3.32) can be seen as (3.33) with $\phi(\sigma) = \sigma^c \in \mathcal{F}_C$, $C = b^c$, $b > \kappa^2$ while for $c > 1$ we can write $\phi(\sigma) = \vartheta(\sigma)\psi(\sigma)$ where $\vartheta(\sigma) = \sigma^p / C_1$ and $\psi(\sigma) = C_1 \sigma^{c-p} \in \mathcal{F}_C$, $C = C_1 b^{c-p}$, $b > \kappa^2$, $C_1 = p\kappa^{2(p-1)}$ and $p = [c]$ is an integer part of c. It is clear that the Lipschitz constant for such*

a $\vartheta(\sigma)$ is equal to 1. At the same time, source conditions (3.33) with $\phi \in \mathcal{F}$ cover all types of smoothness studied so far in Regularization Theory. For example $\psi(\sigma) = \sigma^p \log^{-c} 1/\sigma$ with $p = 0, 1, \ldots$, $c \in [0, 1]$ can be split in a Lipschitz part $\vartheta(\sigma) = \sigma^p$ and an operator monotone part $\psi(\sigma) = \log^{-c} 1/\sigma$

3.6.1 Qualification and Source Condition

We end this section discussing the important interplay between qualification and a source condition. To this aim we need the following definition from [MP03].

Definition 6. *We say that the qualification ν_0 covers ϕ, if there is $c > 0$ such that*

$$c\frac{\lambda^{\nu_0}}{\phi(\lambda)} \leq \inf_{\lambda \leq \sigma \leq \kappa^2} \frac{\sigma^{\nu_0}}{\phi(\sigma)} \tag{3.36}$$

where $0 < \lambda \leq \kappa^2$.

The following important result is a restatement of proposition 3 in [MP03]. Its importance lies in the fact we can see a relation between condition (3.23) in definition 5 and the notion of source condition.

Proposition 3. *Let ϕ be a non decreasing index function and let g_λ be a regularization with qualification which covers ϕ. Then the following inequality holds true*

$$\sup_{0<\sigma\leq\kappa^2} |1 - g_\lambda(\sigma)\sigma|\phi(\sigma) \leq c_g\phi(\lambda), \qquad c_g = \frac{\gamma_\nu}{c},$$

where c is a constant from (3.36).

Remark 6. *The index functions $\phi \in \mathcal{F}$ are covered by regularization with infinite qualification such as spectral cut-off or Landweber iteration. Moreover, from theorem 3 above it follows that the index functions $\phi \in \mathcal{F}_C$ are covered by the qualification of Tikhonov regularization. Note also that if the function $\sigma \to \sigma^\nu/\phi(\sigma)$ is increasing then (3.36) is certainly satisfied with $c = 1$.*

3.7 Discussion and Previous Work

In this section we summarize the main points of our analysis and discuss some connections with previous work on the subject.

Indeed the connection between learning and inverse problems goes beyond analogies in the considered algorithms and we can see similarities and differences between the two theories. In particular we note that:

- Learning with the square loss in RKH spaces is an ill-posed inverse problem. The problem of approximating the regression function in the population case correspond to the problem of inverting a linear embedding equation. On the other hand the sample case induces a linear inverse problem which can be seen as a perturbation of the population problem.

- Such a formulation allows us to adapt a large class of regularization algorithms to the needs of learning. We obtain a class of kernel methods which are easy to implement and can all be described in a unified general framework. In the next chapter we will see that for all these algorithms we can deduce error bounds which highlight different properties of each algorithm.

- Moreover it turns out that we can use the same prior assumptions which are usually considered in inverse problems. Restriction on the decay of the Fourier coefficients of the target function is described in an elegant way in the context of variable Hilbert Scale.

On the other hand we can note some differences. First, rather than the reconstruction error, the residual of the solution is important in Learning Theory since it has an interpretation in terms of expected error. Second, we have to define new perturbation measures. The main reason for this is that the discretization procedure in learning is stochastic and cannot be controlled. This in particular explain the different kind of convergence considered in the two theories: with respect to the number of samples in learning and with respect to the noise in inverse problems.

Finally we discuss connection with previous work on learning, regularization and inverse problems. The main inspiration for our analysis are the seminal works [PG92, GJP95] which opened the way to the interpretation of many learning algorithms in the perspective of regularization (see [EPP00] and reference therein) and highlighted the analogy with inverse problems. Some recent works further developed this point of view extending the notion of stability to include a larger class of algorithms not necessarily related to the square loss function (see [MNPR04, PRMN04] or [RMP05] references therein). With this new definition of stability we can seen again a crucial interplay between well-posedeness and consistency. This latter point of view is also indebted to the work of [BE02] where the relation between stability and some generalization properties is considered.
From a more formal point of view, connections between function approximation problems, regression and inverse problems have been studied in a slightly different context to the one of learning. In the following we recall some of them and give pointers to further readings. If we consider a function approximation problem from fixed grid of input points where the outputs are generated by some target function and are corrupted with some deterministic noise, then it is well known that this is a standard inverse problem [BDMP88]. The error analysis is in this case is mostly dealing with the reconstruction error and stability to

the noise. In statistics usually fixed design model have been considered were again the input points are fixed and the output is generated by some target function but is now corrupted by some stochastic noise. In this case usually mean square error is considered as the error measure which is studied through an analysis in expectation. Many results exists along this line which are extremely similar to those presented here (see for example [LL04, BMR06, Kur04] which are also a good source of references). Often results for fixed design model immediately yields results for models where the inputs are sampled according to a probability distribution (random design) so that the difference between random and fixed design is often not stressed. Nonetheless it is important to note that this difference is negligible if we are interested into results in expectation but becomes crucial as we look for exponential tail inequalities. In particular in the latter case one has to take care explicitly of a variance term which is due to random sampling, (see next chapter).
For this reason previous works attempting a connections between inverse problems and statistical inference problems cannot be applied to the learning setting.

Finally mention some works which considered algorithms inspired by inverse problems. The main examples are algorithms based on Tikhonov regularization which are connected to large margin [Vap98] kernel methods such as Support Vector Machines. More recently [OC04] considered various iterative methods and showed some partial theoretical results. As we already noted boosting algorithm, called $L2$ boost in [BY02] can be seen as a special case of Landweber iteration using splines. From a more general point of view regularization is recognized to be a crucial aspect to avoid over-fitting but the word regularization is used in a broad sense so that essentially all the learning algorithms implement some kind of regularization principle. In our analysis we tried to give a mathematical definition of regularization which allows to see the main properties of the considered algorithms.

Chapter 4

Error Estimates for Regularization with Square Loss

In this chapter we collect all the results concerning theoretical properties of the class of algorithms we presented in the previous chapter. In particular we consider the following issues:

- consistency;
- definition of suitable parameter choice (data independent and data dependent);
- determining the main factors determining such a choice;
- the performance (in terms of error bounds) for the algorithms provided with such a choice;
- highlighting the different properties of the various algorithms.

Before starting our analysis we try to comment on the outcome of our study. First we note that we have different results depending on the range of prior we consider. In fact in our analysis we have to treat separately the case when the best in the model exists and when it does not exist.
In any case we can prove consistency for all the algorithms presented in previous section both for regression and classification. As we previously discussed the rate of convergences depends on suitable assumption on the target function.

If the best in the model exists we get better results matching existing results and extending them: we consider more general noise assumptions and more general a priori assumptions on the target function and especially we give error bounds for new learning algorithms. In

fact the algorithms we introduced in previous chapter can be treated a unified framework which highlights different theoretical properties.
If the best in the model does not exist (note that this case is not treated in inverse problems) things get more difficult and though we can still work in a unified framework better results are obtained for certain algorithms (Tikhonov and Landweber regularization) using ad hoc proofs.
As for the parameter choice in the main consistency results we consider data-independent parameter choice which depends on the a priori assumptions on the problem. Since the latter are usually not known, we propose a data-dependent parameter choice which is adaptive to the unknown prior assumptions when we measure convergence in the RKH space.

The plan of the chapter is as follows: In section 4.1 we give some preliminary considerations. In section 4.2 we discuss error estimates for the case when the best in the model exists and section 4.3 we discuss error estimates for the case when the best in the model does not exist. In section 4.4 we discuss an a posteriori parameter choice which is adaptive in the RKH space. In section 4.5 we discuss implications of our error analysis in a classification setting. Finally in section 4.6 we add some comments and discuss open problems.

4.1 Preliminaries

In this section we gives some preliminaries comments.

We have seen that working with the square loss we have a natural interpretation of the excess error in term of approximation of the regression function with respect to the ρ-norm,

$$\mathcal{E}(f) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) = \|f - Pf_\rho\|_\rho^2,$$

where as usual $P : \mathcal{H} \to L^2(X, \rho_X)$ is the projection on the closure of $\mathcal{H}$ in $L^2(X, \rho_X)$.
Note that for $f \in \mathcal{H}$, if the context allows no confusion, we write $\|f\|_\rho$ in place of $\|I_K f\|_\rho$.
In particular, if $f_\mathcal{H}^\dagger$ exists and $f_\mathbf{z} \in \mathcal{H}$ is an estimator then we consider bounds of the form

$$\Pr\left(\left\| f_\mathbf{z} - f_\mathcal{H}^\dagger \right\|_\rho \geq \varepsilon \right) \leq \eta(\varepsilon, n),$$

for $n \in \mathbb{N}$ and $\varepsilon > 0$. If existence of $f_\mathcal{H}^\dagger$ is not ensured we replace the above inequality with

$$\Pr\left(\|f_\mathbf{z} - Pf_\rho\|_\rho \geq \varepsilon \right) \leq \eta(\varepsilon, n),$$

for $n \in \mathbb{N}$ and $\varepsilon > 0$.

Remark 7. *Note that as we provide bounds and rate for the deviation in the ρ-norm we always have to take the square to go back to bounds and rate for the expected risk.*

As a byproduct of our approach, we easily get error estimates measured in the $\mathcal{H}$-norm. This can be interesting since convergence in $\mathcal{H}$-norm implies point-wise convergence and moreover choosing different kernels we might get convergence in different norms, for example Sobolev norms (see discussion in [SZ05]).

In the following we study the family of algorithms defined by

$$f_{\mathbf{z}}^{\lambda} = g_{\lambda}(T_{\mathbf{x}})S_{\mathbf{x}}\mathbf{y} \tag{4.1}$$

where g_λ satisfies definition 5. Under suitable a priori conditions we show that we can choose $\lambda = \lambda(n)$ in such a way that $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda(n)}$ satisfies probabilistic inequalities like the ones above.

Before starting our analysis we recall that lower rates are available when more information is known on the structure of the RKH space. If $f_\rho \in \Omega_{r,R}$, $r > 1/2$ with $\Omega_{r,R}$ given in (3.31) and the eigenvalues σ_i^2 of the integral operator L_K satisfies $\sigma_i^2 \propto i^{-b}$, $b > 1$ then the following lower rate was proved in [CDV05b]

$$\lim_{A\to 0} \liminf_{n\to\infty} \inf_{f_{\mathbf{z}}} \sup_{\rho\in\mathcal{M}(\Omega_{r,R})} \Pr\left(\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^{\dagger} \right\|_{\rho} \leq A n^{-\frac{rb}{2rb+1}} \right) = 1, \tag{4.2}$$

recalling that $\mathcal{M}(\Omega_{r,R})$ is the set of Borel probability measures such that the regression function is in $\Omega_{r,R}$. The above result is a benchmark to compare our results. Even if it does not directly apply if we drop the assumption on the decay of the eigenvalues, it suggests that the order $n^{-\frac{r}{2r+1}}$ should be optimal in this case.

4.2 Regularization when the Best in the Model Exists

Throughout this section we make the following assumptions.

- **Best in the model**. The following condition holds

  $$Pf_\rho \in \mathcal{H} \tag{4.3}$$

  so that $f_{\mathcal{H}}^{\dagger}$ exists. Moreover we assume that (3.33) holds.

- **Noise assumption**. There exist M, Σ such that the probability measure ρ satisfies the following conditions:

  $$\int_Y \left(e^{\frac{|y - f_{\mathcal{H}}^{\dagger}(x)|}{M}} - \frac{|y - f_{\mathcal{H}}^{\dagger}(x)|}{M} - 1 \right) d\rho(y|x) \leq \frac{\Sigma^2}{2M^2}, \qquad \text{for } \rho_X\text{-almost all } x \in X \tag{4.4}$$

and moreover

$$\int_{X\times Y} y^2 d\rho(y,x) \leq \infty. \tag{4.5}$$

- **Kernel assumption**. The kernel satisfies assumption 1 in particular we recall that $\kappa = \sup_{x\in X}\sqrt{K(x,x)} < \infty$.

Our approach develops in two steps: first we study error bounds for a fixed value of λ, second we define an a priori parameter choice optimizing the obtained bound.
The following result provides us with error estimates for a fixed value of the regularization parameter λ.

Theorem 4. *Let $0 < \lambda \leq 1$. We let $f_{\mathbf{z}}^{\lambda}$ as in (4.1), where g_λ satisfies definition 5. Assume that conditions (4.3), (4.4) and (4.5) hold. Moreover assume $f_{\mathcal{H}}^{\dagger} \in \Omega_{\phi,R}$ and that the regularization has a qualification which covers $\phi(\sigma)\sqrt{\sigma}$. If*

$$\lambda \geq \frac{1}{\sqrt{n}} 2\sqrt{2}\kappa^2 \log\frac{4}{\eta} \tag{4.6}$$

for $0 < \eta < 1$, $n \in \mathbb{N}$ then with probability at least $1-\eta$

$$\left\| f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}^{\dagger} \right\|_{\rho} \leq (C_1\phi(\lambda)\sqrt{\lambda} + C_2\frac{1}{\sqrt{\lambda n}}) \log\frac{4}{\eta}, \tag{4.7}$$

where $C_1 = 2(1+c_\psi)c_g R$ and $C_2 = ((1+c_g)(D+1)CR2\sqrt{2}\kappa^2 + (\sqrt{DB}+B)(\kappa\Sigma + \frac{M}{\sqrt{2}\kappa}))$.

Moreover with probability at least $1-\eta$

$$\left\| f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}^{\dagger} \right\|_{\mathcal{H}} \leq (C_3\phi(\lambda) + C_4\frac{1}{\lambda\sqrt{n}}) \log\frac{4}{\eta}, \tag{4.8}$$

where $C_3 = (1+c_\psi)c_g R$ and $C_4 = ((D+1)CR2\sqrt{2}\kappa^2 + B(\kappa\Sigma + \frac{M}{\sqrt{2}\kappa}))$.

We postpone the proof to the next section. We give some comments on the assumptions in the above theorem. Condition (4.6) has been considered in [SZ05] and [CDV05b, CDV05a] and is not really restrictive. In fact, it is interesting to note that for index function $\phi \in \mathcal{F}$, the above theorem (with larger values of the constants) is still valid under weaker condition $\lambda \geq (c\kappa^2\sqrt{n})^{-1}$ which is not restrictive at all because is satisfied for the best a priori choice of the regularization parameter (see theorem 5 below) balancing the values of the terms in the estimates (4.7) and (4.8). Moreover the condition $\lambda \leq 1$ is considered only to simplify the results and can be replaced by $\lambda \leq a$ for some positive constant a that would eventually appear in the bound.

As for the bounds in the above theorem we add the following remarks.

Remark 8. *If $f_\rho \in \mathcal{H}$ clearly we can replace $f_{\mathcal{H}}^\dagger$ with f_ρ since $\mathcal{E}(f_{\mathcal{H}}^\dagger) = \inf_{f\in\mathcal{H}} \mathcal{E}(f) = \mathcal{E}(f_\rho)$.*

Remark 9. *The main drawback in the above result is that we have to assume the existence of $f_{\mathcal{H}}^\dagger$. Though this assumption is necessary to study result in the $\mathcal{H}$-norm it can be relaxed when looking for bounds in $L^2(X, \rho_X)$ (see discussion in section 3.2). In section 4.3 we discuss in detail the case when $f_{\mathcal{H}}^\dagger$ does not exist.*

Remark 10. *Inspecting the proof of the above theorem we see that the family of good training sets such that the bounds hold with high probability do not depend on the value of the regularization parameter. This turns out to be useful to define a data driven strategy for the choice of λ (see section 4.4).*

From the above results we can immediately derive a data independent (a priori) parameter choice $\lambda_n = \lambda(n)$. The following theorem shows the error bounds obtained providing the one parameter family of algorithms in (4.1) with such a regularization parameter choice.

Theorem 5. *We let $\Theta(\lambda) = \phi(\lambda)\lambda$. Under the same assumptions of theorem 4 we choose*

$$\lambda_n = \Theta^{-1}(n^{-\frac{1}{2}}) \tag{4.9}$$

and let $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$. Then for $0 < \eta < 1$ and $n \in \mathbb{N}$ such that

$$\Theta^{-1}(n^{-\frac{1}{2}})n^{-\frac{1}{2}} \geq 2\sqrt{2}\kappa^2 \log\frac{4}{\eta} \tag{4.10}$$

the following bound holds with probability at least $1-\eta$

$$\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^\dagger \right\|_\rho \leq (C_1 + C_2)\phi(\Theta^{-1}(n^{-\frac{1}{2}}))\sqrt{\Theta^{-1}(n^{-\frac{1}{2}})}\log\frac{4}{\eta},$$

with C_1 and C_2 as in theorem 4. Moreover with probability at least $1-\eta$

$$\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^\dagger \right\|_{\mathcal{H}} \leq (C_3 + C_4)\phi(\Theta^{-1}(n^{-\frac{1}{2}}))\log\frac{4}{\eta},$$

with C_3 and C_4 as in theorem 4.

Several corollaries easily follow. The following result considers the stochastic order [vdG00] of convergence with respect to the ρ-norm.

Corollary 3. *Under the same assumptions of theorem 5 if λ_n is chosen according to (4.9) and $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$ then*

$$\lim_{A\to\infty} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}(\Omega_{\phi,R})} \Pr\left(\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^\dagger \right\|_\rho > A a_n \right) = 0$$

for $a_n = \phi(\Theta^{-1}(n^{-\frac{1}{2}}))\sqrt{\Theta^{-1}(n^{-\frac{1}{2}})}$, $A \in \mathbb{R}^+$.

Up-to our knowledge there are no minimax lower bounds for the class of priors considered here. At the beginning of this section we recalled the lower rates (4.2) obtained for $\rho \in \mathcal{M}(\Omega_{r,R})$, that is Hölder source condition, and for the eigenvalues of T has a polynomial decay $\sigma_i^2 \propto i^{-b}$, $b > 1$. In this case lower rate $a_n = n^{-\frac{rb}{2rb+1}}$, $1/2 < r$ are shown to be optimal. As it can be seen from next corollary we share the same dependence on the smoothness index r though we do not exploit the information on the behavior of the spectrum of T.

The following result considers the case of Hölder source conditions, that is the case when condition (3.33) reduces to (3.32). Recalling the equivalence between (3.31) an (3.32) we state the following result considering $\nu = r - 1/2$ to have an easier comparison with previous results.

Corollary 4. *Under the same assumption of theorem 5 let $\phi(\sigma) = \sigma^\nu$, $\nu = r - 1/2$. Now choose λ_n as in (4.9) and let $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$. Then for $0 < \eta \leq 1$ and*

$$n > \left(2\sqrt{2}\kappa^2 \log \frac{4}{\eta}\right)^{\frac{4r+2}{2r+3}} \tag{4.11}$$

the following bounds hold with probability at least $1 - \eta$

$$\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^\dagger \right\|_\rho \leq (C_1 + C_2) n^{-\frac{r}{2r+1}} \log \frac{4}{\eta},$$

with C_1 and C_2 as in theorem4 and

$$\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^\dagger \right\|_{\mathcal{H}} \leq (C_3 + C_4) n^{-\frac{r-1/2}{2r+1}} \log \frac{4}{\eta},$$

with C_3 and C_4 as in theorem 4.

Remark 11. *Clearly if in place of $Pf_\rho \in \Omega_{r,R}$ we take $f_\rho \in \Omega_{r,R}$ with $r > 1/2$ then $f_\rho \in \mathcal{H}$ and we can replace $f_{\mathcal{H}}^\dagger$ with f_ρ since $\inf_{f \in \mathcal{H}} \mathcal{E}(f) = \mathcal{E}(f_\rho)$.*

In particular we discuss the bounds corresponding to some of the examples of regularization algorithms discussed in Section 3.3.2 and for the sake of clarity we restrict ourselves to polynomial source condition and $\mathcal{H}$ dense.

Remark 12. *In the considered range of prior ($r > 1/2$) the above results match those obtained in [SZ05] for Tikhonov regularization possibly with worse constants. We observe that this kind of regularization suffers from a saturation effect and the results no longer improve after a certain regularity level, $r = 1$ (or $r = 3/2$ for the $\mathcal{H}$-norm) is reached. This is a well known fact in the theory of inverse problems.*

Remark 13. *In the considered range of prior ($r > 1/2$) the above results improve on those obtained in [YRC05] for gradient descent learning (see section 4.3.3). Moreover as pointed out in [YRC05] such an algorithm does not suffer from saturation and the rate can be extremely good if the regression function is regular enough (that is if r is big enough) though the constant gets worse.*

Remark 14. *The spectral cut-off regularization does not suffer from the saturation phenomenon and moreover the constant does not change with the regularity of the solution, allowing extremely good theoretical properties. Note that such an algorithm is computationally feasible if one can compute the SVD of the kernel matrix $\mathbf{K}$.*

Remark 15. *The semiiterative methods though suffering from a saturation effect may have some advantage on Landweber iteration from the computational point of view. In fact recalling that we can identify $\lambda = t^{-2}$ it is easy to see that they require the square root of the number of iterations required by Landweber iteration to get the same convergence rate.*

4.2.1 Proofs

We now give the proofs of the results presented in the previous section. To this aim we make use of the following lemma which is a straightforward corollary of proposition 2.

Lemma 2. *We assume that the kernel is bounded and assumption (4.4) holds. For $0 < \eta \leq 1$ and $n \in \mathbb{N}$ if we let*

$$G = \{\mathbf{z} \in Z^n : \left\| T_{\mathbf{x}} f_{\mathcal{H}}^{\dagger} - S_{\mathbf{x}}^{*}\mathbf{y} \right\|_{\mathcal{H}} \leq \delta_1, \quad \|T - T_{\mathbf{x}}\| \leq \delta_2\},$$

with

$$\begin{aligned} \delta_1 := \delta_1(n,\eta) &= 2(\frac{\kappa M}{n} + \frac{\kappa\Sigma}{\sqrt{n}}) \log\frac{4}{\eta} \\ \delta_2 := \delta_2(n,\eta) &= \frac{2\sqrt{2}\kappa^2}{\sqrt{n}} \log\frac{4}{\eta}, \end{aligned}$$

then

$$\Pr(G) \geq 1 - \eta.$$

We are now ready to prove theorem 4.

Proof of theorem 4. Let

$$r_\lambda(\sigma) = 1 - \sigma g_\lambda(\sigma). \tag{4.12}$$

and δ_1, δ_2, G as in lemma 2. Then from this lemma we know that

$$\Pr(G) \geq 1 - \eta. \tag{4.13}$$

Moreover we note that condition (3.21) immediately yields

$$\sup_{0<\sigma\leq\kappa^2} |1 - g_\lambda(\sigma)\sigma| \leq D + 1 := \gamma, \tag{4.14}$$

where we simply used the triangle inequality.

We consider the following decomposition into two terms

$$\begin{aligned} f^\dagger_{\mathcal{H}} - f^\lambda_{\mathbf{z}} &= f^\dagger_{\mathcal{H}} - g_\lambda(T_{\mathbf{x}})S^*_{\mathbf{x}}\mathbf{y} \\ &= (f^\dagger_{\mathcal{H}} - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f^\dagger_{\mathcal{H}}) + (g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f^\dagger_{\mathcal{H}} - g_\lambda(T_{\mathbf{x}})S^*_{\mathbf{x}}\mathbf{y}). \end{aligned} \tag{4.15}$$

The idea is then to separately bound each term both in the norm in $\mathcal{H}$ and in $L^2(X, \rho_X)$.

We start dealing with the first term. Using (3.33) and (4.12) we can write

$$\begin{aligned} f^\dagger_{\mathcal{H}} - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f^\dagger_{\mathcal{H}} &= (I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})\phi(T)v \\ &= r_\lambda(T_{\mathbf{x}})\phi(T_{\mathbf{x}})v + r_\lambda(T_{\mathbf{x}})(\phi(T) - \phi(T_{\mathbf{x}}))v \\ &= r_\lambda(T_{\mathbf{x}})\phi(T_{\mathbf{x}})v + r_\lambda(T_{\mathbf{x}})\vartheta(T_{\mathbf{x}})(\psi(T) - \psi(T_{\mathbf{x}}))v \\ &\quad + r_\lambda(T_{\mathbf{x}})(\vartheta(T) - \vartheta(T_{\mathbf{x}}))\psi(T)v. \end{aligned} \tag{4.16}$$

When considering the norm in $\mathcal{H}$ we know that proposition 3 applies since ϕ (as well as ϑ) is covered by the qualification of g_λ.

Then we can use (4.14), (3.23), (3.33) and theorem 3 to get the bound

$$\left\| f^\dagger_{\mathcal{H}} - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f^\dagger_{\mathcal{H}} \right\|_{\mathcal{H}} \leq c_g R\phi(\lambda) + c_g c_\psi R\vartheta(\lambda)\psi(\|T - T_{\mathbf{x}}\|) + \gamma C R \|T - T_{\mathbf{x}}\|$$

where C is given in (3.35). For $\mathbf{z} \in G$ we have

$$\left\| f^\dagger_{\mathcal{H}} - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f^\dagger_{\mathcal{H}} \right\|_{\mathcal{H}} \leq (1 + c_\psi)c_g R\phi(\lambda) + \gamma C R\delta_2 \tag{4.17}$$

where we used (4.6) to have $\vartheta(\lambda)\psi(\|T - T_{\mathbf{x}}\|) \leq \vartheta(\lambda)\psi(\delta_2) \leq \vartheta(\lambda)\psi(\lambda) = \phi(\lambda)$. Some more reasoning is needed to get the bound in $L^2(X, \rho_X)$. To this aim in place of (4.16) we consider

$$\sqrt{T}(f^\dagger_{\mathcal{H}} - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f^\dagger_{\mathcal{H}}) = (\sqrt{T} - \sqrt{T_{\mathbf{x}}})(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})f^\dagger_{\mathcal{H}} + \sqrt{T_{\mathbf{x}}}(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})f^\dagger_{\mathcal{H}}. \tag{4.18}$$

The first addend is easy to bound since from condition (4.6) and operator monotonicity of $\psi(\sigma) = \sqrt{\sigma}$ we get

$$\left\| \sqrt{T} - \sqrt{T_{\mathbf{x}}} \right\| \leq \sqrt{\|T - T_{\mathbf{x}}\|} \leq \sqrt{\delta_2} \leq \sqrt{\lambda}. \tag{4.19}$$

for $\mathbf{z} \in G$. Then from the above inequality and from (4.17) we get

$$\left\|(\sqrt{T} - \sqrt{T_{\mathbf{x}}})(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})f_{\mathcal{H}}^\dagger\right\|_{\mathcal{H}} \leq (1 + c_\psi)c_g R\phi(\lambda)\sqrt{\lambda} + \gamma C R\sqrt{\lambda}\delta_2. \tag{4.20}$$

On the other hand the second addend can be furtherly decomposed using (3.33)

$$\begin{aligned}\sqrt{T_{\mathbf{x}}}(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})\phi(T)v &= \sqrt{T_{\mathbf{x}}}r_\lambda(T_{\mathbf{x}})\phi(T_{\mathbf{x}})v \\ &\quad +\sqrt{T_{\mathbf{x}}}r_\lambda(T_{\mathbf{x}})\vartheta(T_{\mathbf{x}})(\psi(T) - \psi(T_{\mathbf{x}}))v \\ &\quad +\sqrt{T_{\mathbf{x}}}r_\lambda(T_{\mathbf{x}})(\vartheta(T) - \vartheta(T_{\mathbf{x}}))\psi(T)v.\end{aligned}$$

Using (4.14), (3.23), (3.33) and theorem 3 we get for $\mathbf{z} \in G$

$$\left\|\sqrt{T_{\mathbf{x}}}(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})f_{\mathcal{H}}^\dagger\right\|_{\mathcal{H}} \leq (1 + c_\psi)c_g R\phi(\lambda)\sqrt{\lambda} + c_g\gamma C R\sqrt{\lambda}\delta_2.$$

where again we used (4.6) to have $\psi(\|T - T_{\mathbf{x}}\|) \leq \psi(\delta_2) \leq \psi(\lambda)$. Now we can put the above inequality and (4.20) together to obtain the following bound in the ρ-norm

$$\left\|\sqrt{T}(f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f_{\mathcal{H}}^\dagger)\right\|_{\mathcal{H}} \leq 2(1 + c_\psi)c_g R\phi(\lambda)\sqrt{\lambda} + (1 + c_g)\gamma C R\sqrt{\lambda}\delta_2. \tag{4.21}$$

We are now ready to consider the second term in (4.28). If we consider the norm in $\mathcal{H}$ we can write

$$g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^*\mathbf{y} = g_\lambda(T_{\mathbf{x}})(T_{\mathbf{x}}f_{\mathcal{H}}^\dagger - S_{\mathbf{x}}^*\mathbf{y})$$

and for $\mathbf{z} \in G$ then condition (3.22) immediately yields

$$\left\|g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^*\mathbf{y}\right\|_{\mathcal{H}} \leq \frac{B}{\lambda}\delta_1. \tag{4.22}$$

Moreover when considering the norm in $L^2(X, \rho_X)$ we simply have

$$\begin{aligned}\sqrt{T}(g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^*\mathbf{y}) &= \sqrt{T_{\mathbf{x}}}g_\lambda(T_{\mathbf{x}})(T_{\mathbf{x}}f_{\mathcal{H}}^\dagger - S_{\mathbf{x}}^*\mathbf{y}) \\ &+ (\sqrt{T} - \sqrt{T_{\mathbf{x}}})g_\lambda(T_{\mathbf{x}})(T_{\mathbf{x}}f_{\mathcal{H}}^\dagger - S_{\mathbf{x}}^*\mathbf{y}).\end{aligned} \tag{4.23}$$

It is easy to show that

$$\left\|\sqrt{T_{\mathbf{x}}}g_\lambda(T_{\mathbf{x}})\right\| \leq \frac{\sqrt{DB}}{\sqrt{\lambda}}$$

in fact we can apply the spectral theorem noting that from Cauchy-Schwarz inequality we have $|\sqrt{\sigma}g_\lambda\sigma|^2 \leq |g_\lambda\sigma||\sigma g_\lambda\sigma|$ where we used (3.21) and (3.22). We can use the definition of δ_1 with the above inequality to bound the first addend in (4.23) and the definition of δ_1

with inequalities (4.19), (3.22) to bound the second addend in (4.23). Then, using (4.6), we have $\sqrt{\delta_2} \leq \sqrt{\lambda}$ so that

$$\left\| \sqrt{T}(g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}} f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^* \mathbf{y}) \right\|_{\mathcal{H}} \leq \frac{\sqrt{DB}}{\sqrt{\lambda}}\delta_1 + \sqrt{\delta_2}\frac{B}{\lambda}\delta_1 \leq \frac{(\sqrt{DB}+B)}{\sqrt{\lambda}}\delta_1. \tag{4.24}$$

for $\mathbf{z} \in G$. We now are in the position to derive the desired bounds.

Recalling (4.13) and (4.28), we can put (4.17) and (4.22) together to get with probability at least $1-\eta$,

$$\left\| f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}^\dagger \right\|_{\mathcal{H}} \leq (1+c_\psi)c_g R\phi(\lambda) + \gamma CR\delta_2 + \frac{B}{\lambda}\delta_1.$$

We can then simplify the above bound. In fact $\delta_2 \leq \delta_2/\lambda$ since $\lambda \leq 1$ so that

$$\gamma CR\delta_2 \leq \log\frac{4}{\eta}\gamma CR2\sqrt{2}\kappa^2\frac{1}{\lambda\sqrt{n}}.$$

Moreover from the explicit expression of δ_1, using (4.6) and $\lambda \leq 1$ it is easy to prove that

$$\frac{B}{\lambda}\delta_1 \leq \log\frac{4}{\eta}B(\kappa\Sigma + \frac{M}{\sqrt{2}\kappa})\frac{1}{\lambda\sqrt{n}}$$

Putting everything together we have (4.8) in fact

$$\left\| f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}^\dagger \right\|_{\mathcal{H}} \leq (C_3\phi(\lambda) + C_4\frac{1}{\lambda\sqrt{n}})\log\frac{4}{\eta}$$

where $C_3 = (1+c_\psi)c_g R)$ and $C_4 = (\gamma CR2\sqrt{2}\kappa^2 + B(\kappa\Sigma + \frac{M}{\sqrt{2}\kappa}))$.

Similarly we can use eq. (2.10) to write

$$\left\| f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}^\dagger \right\|_{\rho} = \left\| \sqrt{T}(f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}^\dagger) \right\|_{\mathcal{H}}$$

and from (4.21) and (4.24) we get with probability at least $1-\eta$

$$\left\| \sqrt{T}(f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}^\dagger) \right\|_{\mathcal{H}} \leq 2(1+c_\psi)c_g R\phi(\lambda)\sqrt{\lambda} + (1+c_g)\gamma CR\sqrt{\lambda}\delta_2 + \frac{(\sqrt{DB}+B)}{\sqrt{\lambda}}\delta_1.$$

which can be furtherly simplified as above to get (4.7). □

Proof of theorem 5. If we choose λ_n as in (4.9) then for n such that (4.10) holds we have that condition (4.6) is verified and we can apply the bounds of theorem 4 to λ_n. The results easily follow noting that the proposed parameter choice is the one balancing the two terms in (4.7) in fact the following equation is verified for $\lambda = \lambda_n$

$$\phi(\lambda)\sqrt{\lambda} = \frac{1}{\sqrt{\lambda n}}$$

($\phi(\lambda) = \lambda^{-1}n^{1/2}$ for the $\mathcal{H}$-norm). □

Proof of corollary 3. We let $A = (C_3 + C_4) \log \frac{4}{\eta}$ and solve with respect to η to get

$$\eta_A = 4e^{-\frac{A}{C_3+C_4}}.$$

Then we know from theorem 5 that for n such that (4.10) holds

$$\Pr\left(\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^{\dagger} \right\|_{\rho} > A\phi(\Theta^{-1}(n^{-\frac{1}{2}})\sqrt{\Theta^{-1}(n^{-\frac{1}{2}})} \right) \leq \eta_A$$

and

$$\limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}(\Omega_{\phi,R})} \Pr\left(\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^{\dagger} \right\|_{\rho} > A\phi(\Theta^{-1}(n^{-\frac{1}{2}})\sqrt{\Theta^{-1}(n^{-\frac{1}{2}})} \right) \leq \eta_A.$$

The theorem is proved since $\eta_A \to 0$ as $A \to \infty$. □

Proof of corollary 4. By a simple computation we have $\lambda_n = \Theta^{-1}(n^{1/2}) = n^{-\frac{1}{2r+1}}$. Moreover condition (4.10) can now be written explicitly as in (4.11). The proof follows plugging the explicit form of ϕ and λ_n in the bounds of theorem 5. □

4.3 Regularization when the Best in the Model does not Exist

Throughout this section we make the following assumptions.

- **Best in the model**. We drop the assumption $Pf_\rho \in \mathcal{H}$ and we consider instead the prior $Pf_\rho \in \Omega_{\phi,R}$ with

$$\Omega_{\phi,R} = \{f \in L^2(X, \rho_X) : f = \phi(L_K)h, \|h\|_\rho \leq R\} \tag{4.25}$$

  which is the analogue of condition (3.33) (when $f_{\mathcal{H}}^{\dagger}$ does not exists) and is the generalization of $Pf_\rho \in \Omega_{r,R}$, see (3.31). We also assume that the index function ϕ satisfies the conditions discussed in chapter 3, section 3.6.

- **Noise assumption**. The output space is bounded, that is

$$Y = [-M, M], \qquad M > 0. \tag{4.26}$$

- **Kernel assumption**. The kernel satisfies assumption (1) in particular we recall that $\kappa = \sup_{x\in X} \sqrt{K(x,x)} < \infty$.

- Moreover we need to furtherly restrict the class of regularization algorithm we consider. In fact we require that the following Lipschitz condition holds

$$|g_\lambda(\sigma) - g_\lambda(\sigma')| \leq \frac{L}{\lambda^\mu}|\sigma - \sigma'|, \qquad \sigma, \sigma' \in [0, \kappa^2] \tag{4.27}$$

where L is a constant independent to λ and μ a positive coefficient. The above condition is quite natural since it ensures stability with respect to perturbations of the covariance operator T and in practice we only have the empirical covariance operator $T_{\mathbf{x}}$ based on the training set. In fact, the Lipschitz property of an operator function can be characterized in terms of property of the corresponding real valued function. Indeed theorem 8.1 in [BS03] ensures that condition 4.27 implies

$$\|g_\lambda(B_1) - g_\lambda(B_2)\| \leq \frac{L}{\lambda^\mu}\|B_1 - B_2\|$$

where B_1, B_2 belongs to the Banach space of normal operators endowed with the uniform norm and have spectrum in $[0, \kappa^2]$. The exponent μ will essentially determine the rate of convergence of each algorithm.

Remark 16 (Lipschitz condition). *The last condition leaves out only the spectral cut-off regularization algorithm and for all the other algorithms it is easy to check that it holds with $\mu = 2$. Moreover the constant L can be explicitly calculated for all the algorithms. As we will see in the following assuming the regularization to be Lipschitz is the main cause of a worsening of the results in the previous section. We can get better results if we focus on some specific algorithm such as Tikhonov or Landweber iteration. At this stage it is still not clear if for general regularization methods we need to enforce such a restriction or we can get rid of it.*

We now sketch the approach we follow and then state the main results of this section. We only discuss error estimates in $L^2(X, \rho_X)$ since for estimates in $\mathcal{H}$ we have to require existence of the best in the model $\mathcal{H}$ and we already discussed such a situation. The idea is to consider the following error decomposition for fixed λ

$$\left\|f_{\mathbf{z}}^\lambda - Pf_\rho\right\|_\rho \leq \left\|f_{\mathbf{z}}^\lambda - f^\lambda\right\|_\rho + \left\|f^\lambda - Pf_\rho\right\|_\rho \tag{4.28}$$

which can be seen as a kind of bias variance decomposition. If we can find bounds for both terms then we can select the parameter optimizing the obtained bound.

Theorem 6 (Bound for General Regularization). *Let $f_{\mathbf{z}}^\lambda = g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}\mathbf{y}$ where g_λ satisfies definition 5 and such that condition (4.27). Assume that conditions (4.26), (1) hold and that $Pf_\rho \in \Omega_{\phi,R}$ with $\Omega_{\phi,R}$ given in (4.25). Assume that the regularization has*

qualification covering ϕ. Then for $0 < \eta \leq 1$, $n \in \mathbb{N}$ and $\lambda > 0$ the following inequality holds with probability at least $1 - \eta$

$$\left\| f_{\mathbf{z}}^{\lambda} - P f_{\rho} \right\| \leq \left(\frac{C_1}{\lambda^{\beta} \sqrt{n}} + c_g R \phi(\lambda) \right) \log \frac{4}{\eta} \tag{4.29}$$

where $\beta = \max\{1/2, \mu\}$, $C_1 = 4\sqrt{2}\kappa M \left(\sqrt{DB} + \kappa^3 L \right)$ and c_g as in theorem 3.

The following corollary is straightforward.

Corollary 5. *Under the same assumption of the above theorem for $\phi(\lambda) = \lambda^r$ we can take*

$$\lambda_n = n^{-\frac{1}{2(r+\beta)}} \tag{4.30}$$

and for $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$ the following inequality holds with probability at least $1 - \eta$

$$\| f_{\mathbf{z}} - P f_{\rho} \| \leq \log \frac{4}{\eta} (C_1 + \gamma_r R) n^{-\frac{r}{2(r+\beta)}}. \tag{4.31}$$

Remark 17. *Recall that for all the considered methods we have $\mu = 2$ (see remark 16) so that the final rate is $n^{-\frac{r}{2(r+2)}}$.*

We postpone the proof of the above theorem and see how it can be improved if we consider specific regularization. The best results for this prior were obtained for Tikhonov regularization in [SZ05]. The following result is a straightforward extension of the estimate in [SZ05] to general source condition.

Theorem 7 (Bound for Tikhonov Regularization). *Let $f_{\mathbf{z}}^{\lambda} = (T_{\mathbf{x}} + \lambda I)^{-1} S_{\mathbf{x}} \mathbf{y}$. Assume that conditions (4.26), (1) hold and that $P f_{\rho} \in \Omega_{\phi,R}$ with $\Omega_{\phi,R}$ given in (4.25). Then for $0 < \eta \leq 1$, $n \in \mathbb{N}$, $\lambda > 0$ such that*

$$\lambda \geq \frac{2\sqrt{2}\kappa^2}{\sqrt{n}} \log \frac{4}{\eta}.$$

the following inequality holds with probability at least $1 - \eta$

$$\left\| f_{\mathbf{z}}^{\lambda} - P f_{\rho} \right\|_{\rho} \leq \left(\frac{6\kappa M}{\sqrt{\lambda n}} + c_g \phi(\lambda) R \right) \log \frac{4}{\eta}.$$

Again we easily get an a priori parameter choice.

Corollary 6. *Under the same assumption of the theorem 7 for $\phi(\lambda) = \lambda^r$, $0 < r < 1/2$ we can take*

$$\lambda_n = \frac{2\sqrt{2}\kappa^2}{\sqrt{n}} \log \frac{4}{\eta}$$

and for $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$ we have with probability at least $1 - \eta$

$$\|f_{\mathbf{z}} - Pf_\rho\|_\rho \leq C_T n^{-\frac{r}{2}} \log \frac{4}{\eta}$$

where $C_T = (6\kappa M + R)(2\sqrt{2}\kappa^2)^r$.

We postpone the proof of the above results to section 4.3.2 and give some remarks.

Remark 18. *As we previously mentioned better results can be obtained assuming more information is available on the RKH space. In [CDV05b] the effective dimension,*

$$\mathcal{N}(\lambda) := \mathrm{Tr}((T + \lambda)^{-1}T)$$

was considered as a capacity measure for Tikhonov regularization. If the eigenvalues $(\sigma_i)_{i=1}^\infty$ of the operator T fullfill

$$\sigma_i = o(i^{-b}), \qquad 1 < b < \infty$$

we get

$$\mathcal{N}(\lambda) \propto \lambda^{-\frac{1}{b}}.$$

The effective dimension is shown to be crucial to define a parameter choice attaining optimal convergence rates [CDV05b]. Chossing λ satisfying

$$\mathcal{N}(\lambda) = n\lambda^{2r}$$

it is possible to get rates matching the lower rates in (4.2). Interestingly the effective dimension was considered in [Zha05] and is related to localized Rademacher averages of balls in a RKH space [Men03].

Remark 19. *From the definition of $\mathcal{N}(\lambda)$ we see that it depends on the kernel and the marginal measure. This suggests that in a semi-supervised setting, where besides the training set we have unlabeled inputs available, we might be able to estimate $\mathcal{N}(\lambda)$ efficiently. In [CRDVV05] we showed that, if enough unlabeled data are available, we can replace the effective dimensional with an empirical effective dimension (based on unlabeled data) and it is still possible to define an a priori parameter choice achieving optimal rates.*

The gradient descent algorithm obtained considering Landweber iteration with variable step-size

$$\tau_i = \frac{1}{\kappa^2(i+1)^\theta}$$

was considered in [YRC05] for Hölder source condition and it was shown that the best performance can be found for the fixed step-size $1/\kappa^2$. We give the analysis for variable step-size in section 4.3.3 and give the following result for fixed step-size.

Theorem 8 (Bound for Landweber Regularization). *Let $f_{\mathbf{z}}^t = \tau \sum_{i=0}^{t-1} (I - \tau T_{\mathbf{x}})^i S_{\mathbf{x}}^* \mathbf{y}$, with $\tau = 1/\kappa^2$. Assume that conditions (4.26), (1) hold and that $Pf_\rho \in \Omega_{\phi,R}$ with $\Omega_{\phi,R}$ given in (4.25). Then for $0 < \eta \leq 1$, $n \in \mathbb{N}$, $\lambda > 0$ the following inequality holds with probability at least $1 - \eta$*

$$\left\| f_{\mathbf{z}}^t - Pf_\rho \right\|_\rho \leq 4\sqrt{2} M \frac{t}{\sqrt{n}} + c_g \phi(t) R \tag{4.32}$$

with c_g as in proposition 3.

The following corollary gives an a priori parameter choice.

Corollary 7. *Under the same assumption of the above theorem for $\phi(\lambda) = \lambda^r$, $0 < r \leq 1/2$ we can take*

$$t_n = \lceil n^{\frac{1}{2(r+1)}} \rceil \tag{4.33}$$

and for $f_{\mathbf{z}} = f_{\mathbf{z}}^{t_n}$ we have with probability at least $1 - \eta$

$$\left\| f_{\mathbf{z}} - Pf_\rho \right\|_\rho \leq C_L n^{-\frac{r}{2(r+1)}} \log \frac{4}{\eta}$$

with $C_L = R + 4\sqrt{2}M$.

Remark 20. *In all the above results if $\mathcal{H}$ is dense in $L^2(X, \rho_X)$ clearly we can replace Pf_ρ with f_ρ.*

Remark 21. *Given the above error estimates we can easily mimic the proof of corollary 3 to get the corresponding stochastic order of convergence.*

We add a few comments on the above results. We note that the Lipschitz condition on the regularization, which is essentially always verified, leads to very simple proofs but gives worse results than those obtained with a special analysis for Landweber iteration or Tikhonov regularization. The latter are the best results available but their proof make use of the specific form of Tikhonov regularized solution and it is not clear how to extend such results to general regularization. Anyway when the best in the model does not exist no lower bounds are known and the quality of the results is difficult to evaluate

4.3.1 Error Analysis for General Regularization

Here we focus on the error decomposition (4.28). We first prove the bound for the approximation error. The following results are obtained by means of minor modification from standard results in inverse problem.

Theorem 9 (Approximation Error). *Let $f^\lambda = g_\lambda(T)I_K^* f_\rho$ with g_λ as in definition 5. If $Pf_\rho \in \Omega_{\phi,R}$ with $\Omega_{\phi,R}$ given in (4.25). If the regularization has qualification covering ϕ, then*

$$\left\| f^\lambda - Pf_\rho \right\|_\rho \leq c_g R\phi(\lambda), \tag{4.34}$$

with c_g as in proposition 3.

Proof. Using the following useful equality

$$g_\lambda(I_K^* I_K)I_K^* = I_K^* g_\lambda(I_K I_K^*), \tag{4.35}$$

we can write

$$\left\| Pf_\rho - I_K f^\lambda \right\|_\rho = \| f_\rho - I_K g_\lambda(I_K^* I_K) I_K^* P f_\rho \|_\rho = \| (I - L_K g_\lambda(L_K))\phi(L_K)h \|_\rho ,$$

where we used the fact that $Pf_\rho \in \Omega_{\phi,R}$. Since the qualification of g_λ covers ϕ then we can apply proposition 3 and Schwarz inequality to get (4.34). □

The following result gives the bound for the sample error.

Theorem 10 (Estimation Error). *Let $0 < \lambda \leq 1$, $f_{\mathbf{z}}^\lambda = g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^*\mathbf{y}$, $f^\lambda = g_\lambda(T)I_K^* f_\rho$ with g_λ as in definition 5 and moreover assume that condition (4.27) holds. Then for $\lambda > 0$, $n \in \mathbb{N}$ and $0 < \eta \leq 1$ the following inequality holds with probability at least $1 - \eta$*

$$\left\| f_{\mathbf{z}} - f^\lambda \right\|_\rho \leq C_1 \frac{1}{\lambda^\beta \sqrt{n}} \log \frac{4}{\eta} \tag{4.36}$$

where $C_1 = 2\sqrt{2}\kappa M \left(\sqrt{DB} + \kappa^3 L \right)$ and $\beta = \max\{1/2, \mu\}$.

The proof of the above result is divided into a functional analytical part and a probabilistic part. For the probabilistic analysis we use the following lemma which is a straightforward corollary of proposition 2.

Lemma 3. *We assume that the kernel is bounded and assumption (4.26) holds. For $0 < \eta \leq 1$ and $n \in \mathbb{N}$ if we let*

$$G = \{\mathbf{z} \in Z^n : \| I_K^* f_\rho - S_{\mathbf{x}}^* \mathbf{y} \|_{\mathcal{H}} \leq \delta_1, \quad \| T - T_{\mathbf{x}} \| \leq \delta_2 \},$$

with

$$\begin{aligned}\delta_1 := \delta_1(n,\eta) &= \frac{2\sqrt{2}\kappa M}{\sqrt{n}}\log\frac{4}{\eta}\\ \delta_2 := \delta_2(n,\eta) &= \frac{2\sqrt{2}\kappa^2}{\sqrt{n}}\log\frac{4}{\eta},\end{aligned}$$

then

$$\Pr(G) \geq 1-\eta.$$

We are now ready to prove theorem 10.

Proof of theorem 10. Recall that the polar decomposition of I_K gives

$$\left\|I_K(f_{\mathbf{z}}^{\lambda} - f^{\lambda})\right\|_{\rho} = \left\|\sqrt{T}(f_{\mathbf{z}}^{\lambda} - f^{\lambda})\right\|_{\mathcal{H}}.$$

Then consider the following decomposition

$$\begin{aligned}\sqrt{T}(f_{\mathbf{z}}^{\lambda} - f^{\lambda}) &= \sqrt{T}(g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^*\mathbf{y} - g_\lambda(T)I_K^* f_\rho)\\ &= \sqrt{T}g_\lambda(T_{\mathbf{x}}) - g_\lambda(T))S_{\mathbf{x}}^*\mathbf{y} + \sqrt{T}g_\lambda(T)(S_{\mathbf{x}}^*\mathbf{y} - I_K^* f_\rho). \qquad (4.37)\end{aligned}$$

We have $\sqrt{T} \leq \kappa$ and it is easy to prove that

$$\left\|\sqrt{T}g_\lambda(T)\right\| \leq \sqrt{\frac{BD}{\lambda}}$$

noting that $|\sqrt{\sigma}g_\lambda(\sigma)|^2 \leq |g_\lambda(\sigma)||\sigma g_\lambda(\sigma)|$ and using conditions (3.21), (3.22). If we take the norm in (4.37) we can use the above estimates and condition (4.27) to obtain the following bound

$$\left\|f_{\mathbf{z}}^{\lambda} - f^{\lambda}\right\|_{\rho} \leq \frac{\kappa^2 ML}{\lambda^{\mu}}\left\|T - T_{\mathbf{x}}\right\| + \frac{\sqrt{DB}}{\sqrt{\lambda}}\left\|S_{\mathbf{x}}^*\mathbf{y} - I_K^* f_\rho\right\|_{\mathcal{H}}.$$

For $\mathbf{z} \in G$ as in lemma 3 we have that

$$\left\|f_{\mathbf{z}}^{\lambda} - f^{\lambda}\right\|_{\rho} \leq \frac{1}{\lambda^{\beta} n} 2\sqrt{2}\kappa M\left(\sqrt{DB} + \kappa^3 L\right)\log\frac{4}{\eta}$$

with $\beta = \max\{1/2, \mu\}$. Finally lemma 3 shows that the above inequality holds with probability at least $1-\eta$ for all $n \in \mathbb{N}$ so that the theorem is proved. □

We note that the condition $\lambda < 1$ is considered only to simplify the results and can be replaced by $\lambda < a$ for some positive constant. Given the above error bounds we can easily prove theorem 6 and its corollary.. In fact the bound in (4.29) is found plugging the estimate (4.34) and (4.36) into (4.28). Moreover the parameter choice (4.30) simply follows setting equal the two terms in (4.29).

4.3.2 Error Analysis for Tikhonov Regularization

For the approximation error we simply apply theorem 9 to get

$$\left\|f^\lambda - Pf_\rho\right\|_\rho \leq c_g \phi(\lambda) R$$

under general source condition and in particular

$$\left\|f^\lambda - Pf_\rho\right\|_\rho \leq \lambda^r R, \quad 0 < r \leq 1$$

under Hölder source condition. For the sample error we recall the following result due to [SZ05] for which we give a slightly simplified proof.

Theorem 11. *Let $f^\lambda = (T+\lambda I)^{-1} I_K f_\rho$ and $f_{\mathbf{z}}^\lambda = (T_{\mathbf{x}} + \lambda I)^{-1} S_{\mathbf{x}}^* \mathbf{y}$. Assume conditions 1, 4.26 hold. For $0 < \eta \leq 1$, $n \in \mathbb{N}$ if*

$$\lambda \geq \frac{2\sqrt{2}\kappa^2}{\sqrt{n}} \log \frac{4}{\eta} \tag{4.38}$$

then the following inequality holds with probability at least $1-\eta$

$$\left\|f_{\mathbf{z}}^\lambda - f^\lambda\right\|_\rho \leq \frac{6\kappa M}{\sqrt{\lambda n}} \log \frac{4}{n}.$$

Proof. Recalling the definition of $f_{\mathbf{z}}^\lambda$ and f^λ we can write

$$\begin{aligned} \left\|f_{\mathbf{z}}^\lambda - f^\lambda\right\|_\rho &= \left\|\sqrt{T}\left[(T_{\mathbf{x}}+\lambda I)^{-1} S_{\mathbf{x}}^* \mathbf{y} - (T_{\mathbf{x}}+\lambda I)^{-1}(T_{\mathbf{x}}+\lambda I) f^\lambda\right]\right\|_{\mathcal{H}} \\ &\leq \left\|\sqrt{T}(T_{\mathbf{x}}+\lambda I)^{-1}\right\| \left\|(S_{\mathbf{x}}^* \mathbf{y} - T_{\mathbf{x}} f^\lambda) - (I_K f_\rho - T f^\lambda)\right\|_{\mathcal{H}} \end{aligned} \tag{4.39}$$

where we used $\lambda f^\lambda = I_K f_\rho - T f^\lambda$ and $\|f\|_\rho = \left\|\sqrt{T} f\right\|_{\mathcal{H}}$, $\forall f \in \mathcal{H}$.
We now bound the two terms in (4.39). First we claim that, for $0 < \eta \leq 1$,$n \in \mathbb{N}$, with confidence $1-\eta$

$$\left\|(S_{\mathbf{x}}^* \mathbf{y} - T_{\mathbf{x}} f^\lambda) - (I_K f_\rho - T f^\lambda)\right\|_{\mathcal{H}} \leq \frac{6\kappa M}{\sqrt{\lambda n}} \log \frac{2}{\eta}. \tag{4.40}$$

Consider $\xi : Z \to \mathcal{H}$, defined by $\xi = (y - f^\lambda(x)) K_x$, where $K_x = K(\cdot, x)$. recalling that $\|(T+\lambda I)^{-1} I_K^*\| \leq 1/\sqrt{\lambda}$ the definition of f^λ we have $\left\|f^\lambda\right\|_\infty \leq \kappa \left\|f^\lambda\right\|_{\mathcal{H}} \leq \frac{\kappa M}{\sqrt{\lambda}}$ and we can use (4.26) to get

$$\|\xi\|_{\mathcal{H}} \leq M\kappa \left(1 + \frac{\kappa}{\sqrt{\lambda}}\right)$$

Moreover from (1) and (2.3) we have

$$\begin{aligned}\mathrm{E}[\|\xi\|_{\mathcal{H}}^2] &\leq \int_{X\times Y}(y-f^{\lambda}(x))^2K(x,x)d\rho(x,y)\leq\kappa^2\mathcal{E}(f^{\lambda})\\ &\leq \kappa^2(\left\|f^{\lambda}-f_{\rho}\right\|_{\rho}^2+\mathcal{E}(f_{\rho}))\end{aligned}$$

Using (4.35) we have $f_\rho - I_K f^\lambda = (I - L_K(L_K + \lambda I)^{-1})f_\rho$ so that $\left\|f^\lambda - f_\rho\right\|_\rho \leq \|f_\rho\|$. From (4.26) we have $\|f_\rho\|_\rho \leq M$, $\mathcal{E}(f_\rho) \leq M^2$ and it follows that

$$\mathrm{E}[\|\xi\|_{\mathcal{H}}^2] \leq 2\kappa^2M^2.$$

A direct application of lemma 1 with $H = M\kappa\left(1+\frac{\kappa}{\sqrt{\lambda}}\right)$ and $\sigma^2 = 2\kappa^2M^2$ gives with probability at least $1-\eta/2$

$$\left\|S_{\mathbf{x}}^*\mathbf{y} - T_{\mathbf{x}}f^{\lambda} + I_Kf_\rho - Tf^{\lambda}\right\|_{\mathcal{H}} \leq \frac{2M\kappa}{n}\left(1+\frac{\kappa}{\sqrt{\lambda}}\right)\log\frac{4}{\eta} + \frac{2\sqrt{2}\kappa M}{\sqrt{n}}\log\frac{4}{\eta}$$

and claim follows using condition (4.38) to simplify the above result.
We now deal with the first term in (4.39). We can write

$$\left\|\sqrt{T}(T_{\mathbf{x}}+\lambda I)^{-1}\right\| \leq \left\|\sqrt{T}-\sqrt{T_{\mathbf{x}}}\right\|\left\|(T_{\mathbf{x}}+\lambda I)^{-1}\right\| + \left\|\sqrt{T_{\mathbf{x}}}(T_{\mathbf{x}}+\lambda I)^{-1}\right\|$$

We note that if we take $\eta = \eta/2$ in (3.12) in proposition 2 then condition (4.52) implies

$$\left\|\sqrt{T}-\sqrt{T_{\mathbf{x}}}\right\| \leq \sqrt{\|T-T_{\mathbf{x}}\|} \leq \sqrt{\lambda}$$

Recalling $|\sqrt{\sigma}g_\lambda\sigma|^2 \leq |g_\lambda\sigma||\sigma g_\lambda\sigma|$ we can use conditions (3.21) and (3.22) we get

$$\left\|\sqrt{T}(T_{\mathbf{x}}+\lambda I)^{-1}\right\| \leq \frac{2}{\sqrt{\lambda}}.$$

The proof follows plugging the above inequality and (4.40) (with $\eta = \eta/2$) into (4.39). □

Given the above error bounds the proof of theorem 7 (and its corollary) is straightforward and we omit it since it is identical to that of theorem 5.

4.3.3 Error Analysis for Landweber Iteration with Variable Step-Size

In this section we study a slightly more general version of Landweber iteration where we allow the relaxation parameter to depend on the considered iteration. To this aim for $\sigma \in \mathbb{R}$, define a polynomial of degree $t-k+1$,

$$\pi_k^t(\sigma) = \begin{cases} \prod_{i=k}^t(1-\tau_i\sigma), & k \leq t;\\ 1, & k > t.\end{cases} \tag{4.41}$$

We now let

$$\tau_i = \frac{1}{\kappa^2(i+1)^\theta}, \qquad 0 \le \theta < 1$$

and define the regularization

$$g_t(\sigma) = \sum_{i=0}^{t-1} \tau_i \pi_{i+1}^{t-1}(\sigma). \tag{4.42}$$

It is easy to see that the above regularization satisfies conditions similar to those in definition 5. In fact it is easy to see that

$$\sup_{0<\sigma\le\kappa^2} |g_t(\sigma)| \;\le\; \sum_{i=0}^{t-1} \tau_i \le \frac{1}{\kappa^2(1-\theta)} t^{1-\theta}, \tag{4.43}$$

$$\sup_{0<\sigma\le\kappa^2} |\sigma g_t(\sigma)| \;\le\; 1. \tag{4.44}$$

Moreover a simple calculation [YRC05] shows that for any $\nu > 0$

$$\sup_{0<\sigma\le\kappa^2} |1-\sigma g_t(\sigma)|\sigma^\nu \le \left(\frac{2\nu\kappa^2}{e}\right)^\nu t^{-\nu(1-\theta)}. \tag{4.45}$$

The last inequality immediately yields the following bound for the approximation error under Hölder source condition.

Theorem 12. *Let $f^t = g_t(T)I_K^* f_\rho$ with g_t as in (4.42). If $Pf_\rho \in \Omega_{r,R}$ with $\Omega_{r,R}$ given in (3.31), then*

$$\left\| f^t - Pf_\rho \right\|_\rho \le \left(\frac{2r\kappa^2}{e}\right)^r t^{-r(1-\theta)},$$

$t \in \mathbb{N}$.

Proof. Using equality (4.35) we can write

$$\left\| Pf_\rho - I_K f^t \right\|_\rho = \left\| Pf_\rho - I_K g_t(I_K^* I_K) I_K^* Pf_\rho \right\|_\rho = \left\| (I - L_K g_t(L_K)) L_K^r h \right\|_\rho,$$

where we used the fact that $Pf_\rho \in \Omega_{r,R}$. The theorem is proved since we can use the spectral theorem and (4.45) to estimate the last term in the above equalities. □

For the sample error we can prove the following result.

Theorem 13. *Let $f^t = g_t(T)I_K^* f_\rho$ and $f_{\mathbf{z}}^t = g_t(T_{\mathbf{x}})S_{\mathbf{x}}^*\mathbf{y}$ with g_t as in (4.42). Then the following inequality holds with probability at least $1-\eta$*

$$\left\| f_{\mathbf{z}}^t - f^t \right\|_\rho \le \frac{4\sqrt{2}M}{\sqrt{n}} \frac{t^{1-\theta}}{1-\theta} \log\frac{4}{n}$$

for $0 < \eta \le 1$, $t, n \in \mathbb{N}$.

Proof. We first give a suitable decomposition of $f_{\mathbf{z}}^t - f^t$. To this aim we write

$$f_{\mathbf{z}}^{t+1} = f_{\mathbf{z}}^t + \tau_i(S_{\mathbf{x}}^*\mathbf{y} - T_{\mathbf{x}}f_{\mathbf{z}}^t) = (1-\tau_i T)f_{\mathbf{z}}^t + \tau_i(S_{\mathbf{x}}^*\mathbf{y} + (T-T_{\mathbf{x}})f_{\mathbf{z}}^t)$$

so that, for $f_{\mathbf{z}}^0 = 0$, we get by induction

$$f_{\mathbf{z}}^t = \sum_{i=0}^{t-1} \tau_i \pi_{i+1}^{t-1}(T)[(T-T_{\mathbf{x}})f_{\mathbf{z}}^i + S_{\mathbf{x}}^*\mathbf{y}].$$

If we subtract f^t to each of side of the above inequality we get

$$f_{\mathbf{z}}^t - f^t = \sum_{i=0}^{t-1} \tau_i \pi_{i+1}^{t-1}(T)[(T-T_{\mathbf{x}})f_{\mathbf{z}}^i + S_{\mathbf{x}}^*\mathbf{y} - I_K^* f_\rho].$$

Now let $\chi_i = (T-T_{\mathbf{x}})f_{\mathbf{z}}^i + S_{\mathbf{x}}^*\mathbf{y} - I_K^* f_\rho$ If we take the squared ρ-norm of the expression we can use $\|f\|_\rho = \left\|\sqrt{T}f\right\|_{\mathcal{H}}$ to write

$$\begin{aligned}
\left\|f_{\mathbf{z}}^t - f^t\right\|_\rho^2 &= \left\langle \sqrt{T}\sum_{i=0}^{t-1}\tau_i\pi_{i+1}^{t-1}(T)\chi_i, \sqrt{T}\sum_{i=0}^{t-1}\tau_i\pi_{i+1}^{t-1}(T)\chi_i \right\rangle_{\mathcal{H}} \\
&\leq \|Tg_t(T)\| \left(\sum_{i=0}^{t-1}\tau_i \left\|\pi_{i+1}^{t-1}(T)\right\|\right) (\sup_{0<i\leq t-1}\|\chi_i\|_{\mathcal{H}})^2 \\
&\leq \frac{1}{\kappa^2(1-\theta)} t^{1-\theta} (\sup_{0<i\leq t-1}\|\chi_i\|_{\mathcal{H}})^2 \qquad (4.46)
\end{aligned}$$

where we used (4.44), the fact that $\left\|\pi_{i+1}^{t-1}(T)\right\| \leq 1$ (as can be seen from (4.41)) and $\sum_{i=0}^{t-1}\tau_i \leq \frac{1}{\kappa^2(1-\theta)}t^{1-\theta}$. Moreover recalling that by polar decomposition $S_{\mathbf{x}}^* = \sqrt{T_{\mathbf{x}}}U_x^*$ and noting that $|\sqrt{\sigma}g_t(\sigma)|^2 = |g_t(\sigma)||\sigma g_t(\sigma)|$ we have from inequalities (4.43), (4.44)

$$\sup_{0<i\leq t-1}\left\|g_i(T_{\mathbf{x}})\sqrt{T_{\mathbf{x}}}U^*\mathbf{y}\right\| \leq \frac{1}{\kappa}\sqrt{\frac{t^{1-\theta}}{1-\theta}}M$$

so that we can write

$$\sup_{0<i\leq t-1}\|\chi_i\|_{\mathcal{H}} \leq \left[\frac{1}{\kappa}\sqrt{\frac{t^{1-\theta}}{1-\theta}}M\,\|(T-T_{\mathbf{x}})\| + \|S_{\mathbf{x}}^*\mathbf{y} - I_K^* f_\rho\|_{\mathcal{H}}\right] \qquad (4.47)$$

For $\mathbf{z} \in G$ with G defined in lemma 3 we can estimate $\|(T-T_{\mathbf{x}})\|$ and $\|S_{\mathbf{x}}^*\mathbf{y} - I_K^* f_\rho\|_{\mathcal{H}}$ explicitly so that using (4.46) and (4.47) we get the following inequality

$$\begin{aligned}
\left\|f_{\mathbf{z}}^t - f^t\right\|_\rho &\leq \frac{1}{\kappa}\sqrt{\frac{t^{1-\theta}}{1-\theta}}\left[\frac{1}{\kappa}\sqrt{\frac{t^{1-\theta}}{1-\theta}}\frac{2\sqrt{2}\kappa^2 M}{\sqrt{n}} + \frac{2\sqrt{2}\kappa M}{\sqrt{n}}\right]\log\frac{4}{n} \\
&\leq \frac{4\sqrt{2}M}{\sqrt{n}}\frac{t^{1-\theta}}{1-\theta}\log\frac{4}{n}
\end{aligned}$$

where we used $\sqrt{t^{1-\theta}/(1-\theta)} \geq 1$. The theorem is proved since lemma 3 ensures that the above inequality holds with probability at least $1-\eta$ for $0 < \eta \leq 1$, $n \in \mathbb{N}$. □

Using the above estimates we easily have the following result.

Corollary 8. *Under the same assumption of theorems 12 and 13, if we let*

$$t_n = \lceil n^{\frac{1}{2(r+1)(1-\theta)}} \rceil$$

and $f_{\mathbf{z}} = f_{\mathbf{z}}^{t_n}$ *then for* $0 < \eta \leq 1$, $n \in \mathbb{N}$, *the following inequality holds with probability at least* $1-\eta$

$$\|f_{\mathbf{z}} - Pf_\rho\|_\rho \leq C_V n^{-\frac{r}{2(r+1)}} \log \frac{4}{\eta}$$

with $C_V = \left(\frac{2r\kappa^2}{e}\right)^r + \frac{4\sqrt{2}M}{1-\theta}$.

Remark 22. *The rate of convergence in the above algorithm is independent to the value of* θ *and the number of iteration is minimized for* $\theta = 0$. *This is an interesting observation since it indicates that step-size in the gradient descent has just an algorithmic meaning and plays no role in regularization.*

The above remark motivates the study of Landweber iteration for fixed step-size in theorem 8 and its corollary.

Proof of theorem 8 and its corollary. For fixed step-size we can easily get estimates for the approximation error under general source condition, replacing $\Omega_{r,R}$ with $\Omega_{\phi,R}$ given in (4.25). In fact we can simply apply theorem 9 to get

$$\left\|f^t - Pf_\rho\right\|_\rho \leq c_g \phi(t^{-1}) R \tag{4.48}$$

with c_g as in proposition 3. For $\phi(t^{-1}) = t^{-r}$, $0 < r < 1/2$ we have $c_g = 1$.

Moreover for $\theta = 0$ the estimate in theorem 13 gives with probability at least $1-\eta$

$$\left\|f_{\mathbf{z}}^t - f^t\right\|_\rho \leq 4\sqrt{2}M \frac{t}{\sqrt{n}} \log \frac{4}{n} \tag{4.49}$$

for $0 < \eta \leq 1$, $n \in \mathbb{N}$, $t \in \mathbb{N}$. The bound (4.32) follows plugging (4.48) and (4.49) into (4.28). Finally the stopping rule in (4.33) follows taking $\phi(t^{-1}) = t^{-r}$, $0 < r < 1/2$ in (4.32) and optimizing w.r.t. to t. □

4.4 Adaptive Regularization in RKH spaces

In the previous sections we discussed how we can choose the regularization parameter λ when the prior assumption on the problem is known. Indeed this is usually not the case and we would prefer to have data driven parameter choices.

In what follows we discuss a Lepskii [Lep90] type strategy to choose λ in the case when the best in the model exists. In particular we consider an approach which has been studied extensively in the context inverse problems (see [SP03] or [BP05] and references therein). Such a choice can be shown to be adaptive to the unknown prior when we measure the error in the norm in the RKH space $\mathcal{H}$. For this reason we have to assume that the best in the model exists.
Though we cannot prove the same result in the ρ-norm, inspecting theorem 5 we can see that the best possible a priori parameter choice is the *same* both for the ρ-norm and the $\mathcal{H}$-norm. This suggests that if have a good parameter choice in the $\mathcal{H}$-norm it might provide a good parameter choice in the ρ-norm.

We assume that the assumptions of section 4.2 hold. We let

$$\lambda_{opt} = \Theta^{-1}\left(n^{-\frac{1}{2}}\right).$$

Recall that this parameter choice is the one which solves the equation

$$\phi(\lambda) = \frac{1}{\sqrt{n}\lambda}$$

induced by the bound in theorem 4. In place of the interval $(0, \kappa^2]$ we now consider only a finite number of values for the regularization parameter. We consider a suitable discretization such that

$$\Lambda_{q,N} = \{\lambda_i : 0 < \lambda_{opt} < \lambda_1 < \cdots < \lambda_N, \quad \lambda_i < q\lambda_{i-1}\}.$$

Example 4. *A way to define the set $\Lambda_{q,N}$ is to consider the geometric series*

$$\lambda_i = \lambda_0 q^i$$

for some initial value λ_0 and with $q > 1$.

The best approximation to λ_{opt} in $\Lambda_{q,N}$ is defined by $\lambda_* = \lambda_*(n)$ such that

$$\lambda_* = \max\{\lambda_i \in \Lambda_{q,N} : \phi(\lambda_i) \leq \frac{1}{\sqrt{n}\lambda_i}\}.$$

It is easy to show that the choosing λ_* we can obtain the same rate as λ_{opt}, nonetheless we still have to know ϕ beforehand (and hence the regularity of the solution) to perform

such a choice. If the smoothness is not known we can define a data dependent parameter choice $\lambda_+ = \lambda_+(n, \mathbf{z})$ by

$$\lambda_+ = \max\{\lambda_i \in \Lambda_{q,N} : \left\| f_{\mathbf{z}}^{\lambda_i} - f_{\mathbf{z}}^{\lambda_j} \right\|_{\mathcal{H}} \leq 4C \log \frac{4}{\eta} \frac{1}{\sqrt{n}\lambda_j}, j = 0, 1, \ldots, i\}.$$

where $C = \max\{C_3, C_4\}$, C_3, C_4 given in theorem 4 and η is the confidence level such that the inequality holds true. The intuition behind this choice is based on the observation that if we take two values $\lambda, \tau \in \Lambda_{q,N}$ such that $\lambda \leq \tau \leq \lambda_*$ then

$$\begin{aligned} \left\| f_{\mathbf{z}}^{\lambda} - f_{\mathbf{z}}^{\tau} \right\|_{\mathcal{H}} &\leq \left\| f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}^{\dagger} \right\|_{\mathcal{H}} + \left\| f_{\mathbf{z}}^{\tau} - f_{\mathcal{H}}^{\dagger} \right\|_{\mathcal{H}} \\ &\leq \log \frac{4}{\eta} (C_3 \phi(\lambda) + C_4 \frac{1}{\sqrt{n}\lambda} + C_3 \phi(\tau) + C_4 \frac{1}{\sqrt{n}\tau}) \\ &\leq 4C \log \frac{4}{\eta} \frac{1}{\sqrt{n}\lambda}. \end{aligned} \tag{4.50}$$

Next theorem shows that in fact such a choice shares the same stochastic order of λ_* and hence of λ_{opt}.

Theorem 14. *Under the same assumptions of theorem 4 we have*

$$\left\| f_{\mathbf{z}}^{\lambda_+} - f_{\mathcal{H}}^{\dagger} \right\|_{\mathcal{H}} \leq 6qC\phi(\Theta^{-1}\left(n^{-\frac{1}{2}}\right)).$$

where $C = \max\{C_3, C_4\}$.

Proof. We previously note that it is easy to prove that $\lambda_* < \lambda_+$, in fact for any $\tau < \lambda_*$ we can repeat the same argument as in (4.50) to get

$$\left\| f_{\mathbf{z}}^{\lambda_*} - f_{\mathbf{z}}^{\tau} \right\|_{\mathcal{H}} \leq 4C \log \frac{4}{\eta} \frac{1}{\sqrt{n}\tau}.$$

Now we can use the definition of λ_+ and λ_* to get

$$\begin{aligned} \left\| f_{\mathbf{z}}^{\lambda_+} - f_{\mathcal{H}}^{\dagger} \right\|_{\mathcal{H}} &\leq \left\| f_{\mathbf{z}}^{\lambda_+} - f_{\mathbf{z}}^{\lambda_*} \right\|_{\mathcal{H}} + \left\| f_{\mathbf{z}}^{\lambda_*} - f_{\mathcal{H}}^{\dagger} \right\|_{\mathcal{H}} \\ &\leq 4C \log \frac{4}{\eta} \frac{1}{\sqrt{n}\lambda_*} + 2C \log \frac{4}{\eta} \frac{1}{\sqrt{n}\lambda_*}. \end{aligned} \tag{4.51}$$

Finally we have to relate λ_* and λ_{opt}. If we let $\lambda_* = \lambda_\ell$ then $\lambda_* = \lambda_\ell \leq \lambda_{opt} \leq \lambda_{\ell+1}$ and since we choose λ from $\Lambda_{q,N}$ we have $\lambda_{opt} \leq q\lambda_*$ that is

$$\frac{1}{\lambda_*} \leq \frac{q}{\lambda_{opt}}. \tag{4.52}$$

If we now recall the definition of λ_{opt} we can use (4.52) and (4.51) to get

$$\left\| f_{\mathbf{z}}^{\lambda_+} - f_{\mathcal{H}}^{\dagger} \right\|_{\mathcal{H}} \leq 6qC \log \frac{4}{\eta} \frac{1}{\sqrt{n}\lambda_{opt}} = 6qC \log \frac{4}{\eta} \phi(\Theta^{-1}(n^{-1/2}))$$

and the theorem is proved. □

We can refine the above method allowing only comparison of solution corresponding to adjacent values of λ. If we now assume that the set $\Lambda_{q,N}$ is defined by the geometric sequence $\lambda_i = \lambda_0 q^i$, with $i = 0, 1, \ldots, N$, then we can define a data dependent choice $\bar{\lambda} = \bar{\lambda}(n, \mathbf{z})$ considering

$$\bar{\lambda} = \max\{\lambda_i \in \Lambda_{q,N} : \left\| f_{\mathbf{z}}^{\lambda_j} - f_{\mathbf{z}}^{\lambda_{j-1}} \right\|_{\mathcal{H}} \leq 4C \log \frac{4}{\eta} \frac{1}{\sqrt{n}\lambda_{j-1}}, j = 0, 1, \ldots, i\}.$$

Again we argue that such value of λ should be sufficiently close to λ_{opt} and next theorems shows that this is indeed the case.

Theorem 15. *Under the same assumptions of theorem (4) we have*

$$\left\| f_{\mathbf{z}}^{\bar{\lambda}} - f_{\mathcal{H}}^{\dagger} \right\|_{\mathcal{H}} \leq C_q \phi(\Theta^{-1}\left(n^{-\frac{1}{2}}\right)).$$

where $C_q = 6\left(\frac{2q-1}{q-1}\right) \max\{C_3, C_4\}$.

Proof. The idea is to observe that we can easily control the distance between the solutions corresponding to λ_* and $\bar{\lambda}$. In fact if we let $\lambda_* = \lambda_\ell$ and $\bar{\lambda} = \lambda_m$ clearly $m > \ell$ and we can use the definition of $\bar{\lambda}$ to write

$$\begin{aligned} \left\| f_{\mathbf{z}}^{\bar{\lambda}} - f_{\mathbf{z}}^{\lambda_*} \right\|_{\mathcal{H}} &\leq \sum_{j=1}^{m-\ell} \left\| f_{\mathbf{z}}^{\lambda_{m-j}} - f_{\mathbf{z}}^{\lambda_{m-j+1}} \right\|_{\mathcal{H}} \leq 4C \log \frac{4}{\eta} \frac{1}{\sqrt{n}} \sum_{j=1}^{m-\ell} \frac{1}{\lambda_{m-j}} \qquad (4.53) \\ &= 4C \log \frac{4}{\eta} \frac{1}{\sqrt{n}\lambda_*} \sum_{j=1}^{m-\ell} \frac{1}{q^{m-j}} \leq 4C \log \frac{4}{\eta} \frac{1}{\sqrt{n}\lambda_*} \frac{q}{q-1} \end{aligned}$$

Then we can consider

$$\begin{aligned} \left\| f_{\mathbf{z}}^{\bar{\lambda}} - f_{\mathcal{H}}^{\dagger} \right\|_{\mathcal{H}} &\leq \left\| f_{\mathbf{z}}^{\lambda_*} - f_{\mathcal{H}}^{\dagger} \right\|_{\mathcal{H}} + \left\| f_{\mathbf{z}}^{\bar{\lambda}} - f_{\mathbf{z}}^{\lambda_*} \right\|_{\mathcal{H}} \qquad (4.54) \\ &\leq 2C \log \frac{4}{\eta} \frac{1}{\sqrt{n}\lambda_*} + 4C \log \frac{4}{\eta} \frac{1}{\sqrt{n}\lambda_*} \frac{q}{q-1} \end{aligned}$$

where we used the definition of λ_*. The theorem is proved using (4.52). □

4.5 Regularization for Binary Classification: Risk Bounds and Bayes Consistency

We discuss the performance of the proposed class of algorithms in the context of binary classification [BBL04b], that is when $Y = \{-1, 1\}$. The problem is that of discriminating the elements of two classes and as usual we can take $\text{sign} f_{\mathbf{z}}$ as our decision rule. In this case some natural error measures can be considered. The *risk* or *misclassification error* is defined as

$$R(f) = \Pr\left(\text{sign} f(x) \neq y\right), \tag{4.55}$$

whose minimizer is the *Bayes rule*

$$f_B(x) = \begin{cases} 1, & \text{if} \quad \rho(1|x) > 1/2; \\ -1, & \text{otherwise.} \end{cases}$$

The latter can be explicitly related to the regression function in fact $f_B = \text{sign} f_\rho$. This can be easily seen recalling that, from the definition of f_ρ, in classification we have

$$f_\rho(x) = 2\rho(1|x) - 1.$$

The goal is then to approximate the Bayes rule and we consider the following quantity to measure the quality of such approximation the *excess risk*

$$R(f_{\mathbf{z}}) - R(f_\rho)$$

and moreover as proposed in [SZ05] we can consider

$$\|\text{sign} f_{\mathbf{z}} - \text{sign} f_\rho\|_\rho .$$

Note that while looking at the excess risk we aim at finding a an estimator whose error is close to that of the Bayes rule f_B; on the other hand in the latter error we aim to recover an estimator whose close to f_B itself (though we put more weight on the points which are most likely to be sampled). For this reason we expect the excess risk to be a weaker error measure and indeed this can be seen from proposition 4 below. To obtain bounds on the above quantities the idea is to relate them to $\|f_{\mathbf{z}} - f_\rho\|_\rho$, that is to find *comparison* results. Though we can obtain some straightforward comparison results, it is interesting to consider some quantity assessing the regularity of the conditional probability. In fact the latter characterizes the noise affecting the classification problem.

To this aim we define the *misclassification set*

$$X_f := \{x \in X \mid \text{sign} f \neq \text{sign} f_\rho\}.$$

The *Tsybakov function* [SZ05] $T_\rho : [0,1] \to [0,1]$ by

$$T_\rho(s) = \rho_X(\{x \in X \ : \ f_\rho(x) \in [-s,s]\}), \tag{4.56}$$

characterizes the probability of level sets of f_ρ. The following *Tsybakov's noise condition* [Tsy04] for some $q \in [0,\infty]$,

$$T_\rho(s) \leq B_q s^q, \quad \forall s \in [0,1], \tag{4.57}$$

characterizes the decay rate of $T_\rho(s)$. In particular when T_ρ vanishes at a neighborhood of 0 (i.e. $T_\rho(s) = 0$ when $s \leq \epsilon$ for some $\epsilon > 0$) we have $q = \infty$.

The following equivalent formulation of the noise condition is useful (see [Tsy04] or [BBL04a]].

Lemma 4. *Tsybakov's condition (4.57) is equivalent* [1]

$$\rho_X(X_f) \leq c_\alpha (R(f) - R(f_\rho))^\alpha, \tag{4.58}$$

where

$$\alpha = \frac{q}{q+1} \in [0,1] \tag{4.59}$$

and $c_\alpha = B_q + 1 \geq 1$.

Given the above premises we can derive several results relating the different error measures introduced before.

Proposition 4. *Let f be some function in $L^2(X, \rho_X)$. The following inequalities hold*

1) $R(f) - R(f_\rho) \leq \|f - f_\rho\|_\rho$

2) If (4.58) hold then $R(f) - R(f_\rho) \leq 4c_\alpha \|f - f_\rho\|_\rho^{2/(2-\alpha)}$

3) $R(f) - R(f_\rho) \leq \frac{1}{2}\|f_\rho\| \|signf - signf_\rho\|_\rho$

4) $\|signf - signf_\rho\|_\rho^2 \leq T(\|f - f_\rho\|_\infty)$

5) If (4.58) hold then $\|signf - signf_\rho\|_\rho \leq 4c_\alpha \|f - f_\rho\|_\rho^{\frac{\alpha}{2-\alpha}}$

Remark 23. *Part 4 was used in [SZ05] by applying bounds on $\|f - f_\rho\|_K$ to estimate $\|f - f_\rho\|_\infty$. Due to the square on the left hand side, this loses a power of $1/2$ in the asymptotic rate. But turning to the weaker norm $\|f - f_\rho\|_\rho$, Part 5 remedies this problem without losing the rate.*

[1] The uniform condition, for all $f \in L^2(X, \rho_X)$, is crucial for the direction as shown in the proof. If we replace it by $f \in \mathcal{H}$, the two conditions are not equivalent. However, the proof of proposition 4 only requires the direction

Using proposition 4 and theorem 5 leads to the following result

Corollary 9. *Assume that $\mathcal{H}$ is dense in $L^2(X, \rho_X)$ and that the same assumptions of theorem 5 hold. Choose λ_n according to (4.9) and let $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$. Then for $0 < \eta < 1$ and n satisfying (4.10) the following bounds hold with probability at least $1 - \eta$*

$$\begin{aligned} R(f_{\mathbf{z}}) - R(f_\rho) &\leq 4c_\alpha \left((C_1 + C_2)\phi(\Theta^{-1}(n^{-\frac{1}{2}})\sqrt{\Theta^{-1}(n^{-\frac{1}{2}})} \log \frac{4}{\eta} \right)^{\frac{2}{2-\alpha}}, \\ \|sign f_{\mathbf{z}} - sign f_\rho\|_\rho &\leq 4c_\alpha \left((C_1 + C_2)\phi(\Theta^{-1}(n^{-\frac{1}{2}})\sqrt{\Theta^{-1}(n^{-\frac{1}{2}})} \log \frac{4}{\eta} \right)^{\frac{\alpha}{2-\alpha}}, \end{aligned}$$

with C_1,C_2,C_3 and C_4 given in theorem 4.

Remark 24 (Fast rates for Bayes consistency). *Corollary 4 shows that for polynomial source conditions this means all the proposed algorithms achieve risk bounds on $R(f_{\mathbf{z}}) - R(f_\rho)$ of order $n^{-\frac{2r}{(2r+1)(2-\alpha)}}$ if n is big enough (satisfying (4.11)). In other words the algorithms we propose are Bayes consistent with fast rates of convergence.*

4.5.1 Proofs

Proof of lemma 4. (4.57) $\Rightarrow$ (4.58). Recalling that

$$R(f) - R(f_\rho) = \int_{X_f} |f_\rho(x)| d\rho_X \tag{4.60}$$

we have the following chains of inequalities

$$\begin{aligned} R(f) - R(f_\rho) &\geq \int_{X_f} |f_\rho(x)| \chi_{|f_\rho(x)|>t} d\rho_X \geq t \int_{X_f} \chi_{|f_\rho(x)|>t} d\rho_X \\ &= t \left[\int_X \chi_{|f_\rho(x)|>t} d\rho_X - \int_{X/X_f} \chi_{|f_\rho(x)|>t} d\rho_X \right] \\ &\geq t \left[(1 - B_q t^q) - \rho_X(X \backslash X_f) \right] = t(\rho_X(X_f) - B_q t^q) \end{aligned}$$

The proof follows taking

$$t = \left(\frac{1}{B_q + 1} \rho_X(X_f) \right)^{1/q}$$

and setting α as in (4.59).

(4.58) $\Rightarrow$ (4.57). Define for $s > 0$,

$$X_s = \{x \in X : |f_\rho(x)| \leq s\}$$

Choose a $f \in L^2(X, \rho_X)$ such that $\text{sign}f = \text{sign}f_\rho$ on $X \backslash X_s$ and otherwise $\text{sign}f \neq \text{sign}f_\rho$, then $X_f = X_s$. Therefore

$$\rho_X(X_f) = \rho_X(X_s) \leq c_\alpha (R(f) - R(f_\rho))^\alpha \leq c_\alpha (\int_{X_s} |f_\rho(x)| d\rho_X)^\alpha \leq c_\alpha t^\alpha \rho_X(X_s)^\alpha = c_\alpha t^\alpha \rho_X(X_f)^\alpha$$

whence $\rho_X(X_f) \leq c_\alpha^{1/(1-\alpha)} t^{\alpha/(1-\alpha)}$ which recovers (4.57) with $q = \alpha/(1-\alpha)$ and $B_q = c_\alpha^{1/(1-\alpha)}$. □

Proof of proposition 4. 1) The proof is straightforward by noting that

$$|f_\rho(x)| \leq |f(x) - f_\rho(x)| \tag{4.61}$$

when $x \in X_f$. In fact from (4.60)

$$R(f) - R(f_\rho) \leq \int_{X_f} |f(x) - f_\rho(x)| \leq \|f - f_\rho\|_\rho$$

2) The inequality is a special case of theorem 10 in [BJM05]. Here we give the proof for completeness. If we further develop (4.60) we get

$$R(f) - R(f_\rho) = \int_{X_f} |f_\rho(x)| \chi_{|f_\rho(x)| \leq t} d\rho_X(x) + \int_{X_f} |f_\rho(x)| \chi_{|f_\rho(x)| > t} d\rho_X(x).$$

where for $|f_\rho(x)| > t$, $|f_\rho(x)| = |f_\rho(x)|^2 / |f_\rho(x) < \frac{1}{t} |f_\rho(x)|^2$. Then by conditions (4.58) and (4.61) we have

$$R(f) - R(f_\rho) \leq t \rho_X(X_f) + \frac{1}{t} \int_{X_f} |f_\rho(x)|_\rho^2 d\rho_X(x) \leq t c_\alpha (R(f) - R(f_\rho))^\alpha + \frac{1}{t} \|f - f_\rho\|_\rho^2.$$

The result follows by taking $t = \frac{1}{2c_\alpha} (R(f) - R(f_\rho))^{1-\alpha}$ and $(4c_\alpha)^{1/(2-\alpha)} \leq 4c_\alpha$ as $\alpha \in [0, 1]$ and $c_\alpha \geq 1$.

3) From (4.60), simply using Schwarz Inequality we have

$$R(f) - R(f_\rho) = \frac{1}{2} \int_X f_\rho(x) (\text{sign}f(x) - \text{sign}f_\rho(x)) d\rho_X(x) \leq \frac{1}{2} \|f_\rho\|_\rho \|\text{sign}f - \text{sign}f_\rho\|_\rho$$

4) See proposition 2 in [SZ05].

5) The proof follows noting that

$$\|\text{sign}f - \text{sign}f_\rho\|_\rho^2 = 4 \rho_X(X_f)$$

and plugging in (4.58) and part 2). □

4.6 Summary of Results and Open Problems

We briefly summarize the main outcome of our analysis. In particular we highlight the following facts:

- We proved finite sample bounds and fast convergence rates (both for regression and classification) for a large class of learning algorithms. In a unified framework we recovered, or improved, the results for existing algorithms and prove similar results for the new algorithms presented in chapter 3. Our analysis allows us to compare advantages and disadvantages of the various algorithms which have different theoretical properties for example saturation properties which are well known in theory of inverse problems are inherited by the the studied class of algorithms.

- At least when the best in the model exists we were able to extend the prior condition previously studied considering regression functions satisfying general source conditions. Moreover following [CDV05b] we were able to relax the condition on the noise dropping the boundedness condition in favor of a much weaker subgaussian condition.

- We proposed a new a posteriori parameter choice which allows to be adaptive to unknown prior with respect to the norm in the RKH space.

- We presented novel proof techniques which are developments of the approach proposed in [DVCR05] and eventually refined in [SZ05, CDV05b, DVRC$^+$05b, SZ04]. Such an approach strongly relies on the linear structure of algorithms induced by the square loss and allows to take advantage of many well known facts in the theory of inverse problems. In particular we showed that, when the error analysis requires a bias-variation decomposition, standard results for inverse problems are easily adapted to get bounds for the bias term. As for the variance term our probabilistic analysis is based on the study of random variables in Hilbert space and in particular to concentration of random operators. Our analysis does not does not make use of covering numbers, Rademacher complexities or other complexity measure. Indeed there are connections between concentration of random operators and uniform convergence results in functional spaces which we do not consider here.

As for the open problems we find the following three directions for further developments. First we have seen that there is a critical range of prior when the best in the model does not exist. The best results so far were obtained for Tikhonov and rely on the special structure of such an algorithm. It would be interesting to better understand the problem in this range of prior extending the results for Tikhonov to more general regularization. Second, our results do consider the effect of the choice of the kernel. Again such an issue was

considered for Tikhonov regularization in [CDV05b] but the proof relies on properties of Tikhonov regularization. It is not clear how to extend such results to general regularization methods. Finally, probably the biggest open problem is the definition of adaptive, data-driven strategy for the choice of the regularization parameter. From the theoretical point of view it is not clear if we can extend the approach we proposed to obtain adaptation in the ρ-norm. Moreover from the practical point of view we expect such strategies to be too conservative since they are based on worst case analysis which leads to pessimistic estimates of the perturbation levels. In fact preliminary numerical simulations indicate that this leads to over-smoothing choices of the regularization parameter.

Chapter 5

Tikhonov Regularization with Convex Loss functions

In this chapter we consider learning with loss functions other than the square loss. We restrict ourselves to admissible loss functions (according to definition 1) which are in particular convex. Moreover we only consider algorithms inspired by Tikhonov regularization, namely regularization networks [EPP00]. We have already seen that this kind of algorithms defines a one parameter family of estimators solving the problems

$$\min_{f\in\mathcal{H}}\{\mathcal{E}_{\mathbf{z}}(f)+\lambda\left\|f\right\|_{\mathcal{H}}^{2}\}, \tag{5.1}$$

where $\lambda > 0$ is the regularization parameter,

$$\mathcal{E}_{\mathbf{z}}(f)=\frac{1}{n}\sum_{i=1}^{n}\ell(y_i,f(x_i))$$

is the empirical error and $\mathcal{H}$ a RKH space. The final estimator $f_{\mathbf{z}}$ is obtained defining a suitable parameter choice $\lambda=\lambda(n,\mathbf{z})$ and letting $f_{\mathbf{z}}=f_{\mathbf{z}}^{\lambda(n,\mathbf{z})}$.

Most of the analysis in this chapter is devoted to a detailed study of the minimization problem (5.1), though in the end we also recall a possible approach to consistency of regularization networks which is reminiscent of the analysis we did for square loss. Before starting our analysis we summarize the main results we present.

As for the properties of the minimization problem, rather than (5.1), we consider the population version of the algorithm which subsumes (5.1) as a special case. Moreover we admit the presence of an unpenalized offset term in the solution. This is customary in Support Vector Machines [Vap98] where the possible solution are hyperplanes $f(x) = wx$

and an unpenalized constant $b \in \mathbb{R}$ is considered accounting for translation invariance of the solution, that is $wx + b$. Here we allow the unpenalized offset to be a function from a possibly infinite dimensional RKH space. Henceforth we consider the problem

$$\min_{(f,g)\mathcal{H}\times\mathcal{B}} \{\mathcal{E}(f+g) + \lambda \|f\|_{\mathcal{H}}^2\}, \tag{5.2}$$

whose minimizer is (f^λ, g^λ). The contribution of our study of property of the minimization of Tikhonov functionals is threefold.

- First we provide a complete characterization of the explicit form of the solution (f^λ, g^λ) by exploiting the convexity assumption on the loss functions. Our result can be interpreted as a quantitative version of the representer theorem holding both for regression and classification and in which explicit care is taken of the offset space $\mathcal{B}$.

- Second, we discuss the role of the offset space $\mathcal{B}$. The starting point of our discussion is the obvious observation that the solution given by problem (5.2) is not the *pair* (f^λ, g^λ) but the *sum* $f^\lambda + g^\lambda$. In other words the natural hypothesis space is the *sum* $\mathcal{H} + \mathcal{B}$ instead of the *product* $\mathcal{H} \times \mathcal{B}$ (which is not even a space of functions from X to $\mathbb{R}$). For arbitrary loss function we prove that problem (5.2) is equivalent to a kernel method defined on $\mathcal{H} + \mathcal{B}$, which is a RKH space, with a penalty term given by a seminorm.

- Third, for sake of completeness, we study the issues of the existence and uniqueness for problem (5.2). When $\mathcal{B}$ is not the empty set, both issues are not trivial. In particular, for $\mathcal{B}$ equal to the set of constants, we prove existence under very reasonable conditions: for example, for classification, one needs at least two examples with different labels. About uniqueness we show that, for strictly convex loss functions, one has uniqueness if and only if the space $\mathcal{B}$ is small enough to be separated by the measure ρ: for example, in the sample case, this last condition means that a function $g \in \mathcal{B}$ is equal to 0 if and only if $g(x_i) = 0$ for all i. For the hinge loss function, which is convex but not strictly convex, we give an *ad hoc* condition in terms of number of support vectors of the two classes.

Our results include and extends previous results (see discussion in section 5.1). As a byproduct of our analysis we show that different algorithms inspired by Support Vector Machines can be cast in a common framework allowing a better understanding of each algorithm.

Finally, we present an approach to consistency of regularization network. One of the main motivation for studying representation results in the population setting is that they proved to be extremely useful for the study of error estimates for the square loss. The main idea is that we can relate the risk of the solution $\mathcal{E}(f_{\mathbf{z}}^\lambda)$ in the sample case to that of the regularized

solution $\mathcal{E}(f^\lambda)$ in the population case comparing the explicit form of $f_{\mathbf{z}}^\lambda$ and f^λ. Here we present a result which is a simplified version of a result given in [CS05]. Moreover for sake of completeness we recall how to derive bounds in classification setting. Again the convex assumption becomes crucial.

The chapter is divided as follows. In section 5.1 we discuss the main contributions and connection with previous work on the properties of the mimization problem describing tikhonov regularization. In section 5.2 we give a new representer theorem. In section 5.3 we discuss the role of an unpenalized off-set term in the solution. In section 5.4 we discussed the problem of existence and uniqueness. In section 5.5 we apply our results in the sample case and discuss in detail the case of Support Vector Machines. In section 5.6 we show that several proposed algorithms can be seen as regularization networks. In section 5.7 we discuss consistency of regularization networks both for regression and classification.

5.1 Properties of Tikhonov Regularization: Discussion on Previous Works

Before developing our analysis we discuss the relevance of our contribution with respect to previous works. Results about the form of the solution of regularization networks are known in the literature as *representer theorems* (if $\mathcal{B}$ is not trivial they are called *semiparametric representer theorems*). The first result in this direction is due to [KW70], see, also, [Wah90], for the square loss function, but the structure of the proof holds for arbitrary loss function as shown by many authors, see, for example, [CO90] and, in the framework of statistical learning, [SHS01] (in this last paper the penalty term can be any strictly increasing function of the norm). This kind of results shows that, if the $\mathcal{H}$ is a RKH space with kernel K, the estimator $f_{\mathbf{z}}^\lambda$ defined by equation (5.1) can be written as

$$f_{\mathbf{z}}^\lambda(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)\,.$$

The above result holds for arbitrary loss function and for a large class of penalty terms. However, in general, the form of the coefficients α_i is unknown. For the square loss function, the form of the coefficients is well known in the context of inverse problems (see chapter 3) and reduces to solve a linear system of equations. For arbitrary differentiable loss functions, the problem was studied by [PG92, Gir98, Wah98] where the coefficients α_i are solution of a system of algebraic equations. This approach cannot be applied to hinge and ϵ-insensitive loss function [Vap98], since they are not differentiable: the form of the coefficients α_i is recovered only through the Lagrangian formulation of the minimization problem, see, for example, [Vap98, CST00]. Recently, [Zha01] gives a quantitative representer theorem in the classification setting that holds for differentiable loss function and [Ste03] extends

this result for arbitrary convex loss function, without using the dual problem. In these papers the form of the coefficients α_i is given in terms of a closed equation involving the subgradient of the loss function. Moreover, they are able to extend the representer theorem to the population case (a study of the solution of Tikhonov regularization in the population case when the square loss is used can be found also in[CS02b]). Our analysis, using techniques similar to those in [Ste03], extends the above result in the following points:

- Our result holds both for regression and classification;
- We provide a general result that holds also when the offset term is considered. The presence of the offset space forces the coefficients α_i to satisfy a system of linear equations;
- We do not assume the input space X and the output space Y to be compact. In particular, for regression we can assume $Y = \mathbb{R}$;
- We provide a simpler proof than the one of [Ste03] using known results in convex analysis.

A discussion of the role of the offset terms can be found in [EPP00] and in [PMR$^+$02] when the space $\mathcal{B}$ reduces to the set of constant functions. The results are close to our theorem 17, but they are proved assuming that the unit constant is in the RKH space and for the sample case. Our results hold true in the population case and for offset term living in arbitrary RKH spaces. The problem of existence and uniqueness is discussed in [Wah98] for the sample case and with differentiable loss functions. For arbitrary ρ the papers by [Ste04, Ste03] study the existence for classification setting with offset space reduced to the constant functions. For the hinge loss and ϵ-insensitive loss, the problem of uniqueness is treated in [BC00, BC03]. Their proof is based on the dual problem and on the Kuhn-Tucker conditions. Our results subsume the cited results as special cases, but are all obtained in the more general population case. In particular our results on uniqueness of SVM solution are similar to those in [BC00, BC03] but do not make use of the dual formulation.

5.2 Explicit form of the regularized solution

In this section we determine the explicit form of the minimizer of the Tikhonov functional introduced in the previous section. We first state the main theorem and comment on the obtained result, then we provide the mathematical proof.

Theorem 16. *Let ρ be a probability measure on $X \times Y$ where X is a Polish space and Y is a closed subset of $\mathbb{R}$. Let ℓ be an admissible loss function with respect to ρ, $p \in [1, +\infty[$.*

Let $\mathcal{H}$ and $\mathcal{B}$ reproducing kernel Hilbert spaces such that the corresponding kernels K and $K_{\mathcal{B}}$ satisfy assumption 1. Define $q =]1, +\infty]$ such that $\frac{1}{q} + \frac{1}{p} = 1$. Let $\lambda > 0$ and $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$, then

$$(f^\lambda, g^\lambda) \in \operatorname*{argmin}_{(f,g)\in\mathcal{H}\times\mathcal{B}} \left\{ \int_{X\times Y} \ell(y, f(x) + g(x)) d\rho(x, y) + \lambda \|f\|^2_{\mathcal{H}} \right\} \tag{5.3}$$

if and only if there is $\alpha \in L^q(Z, \rho)$ satisfying

$$\begin{aligned} \alpha(x, y) &\in (\partial\ell)(y, f^\lambda(x) + g^\lambda(x)) \quad (x, y) \in X \times Y \text{ a.e.} && (5.4)\\ f^\lambda(s) &= -\frac{1}{2\lambda} \int_{X\times Y} K(s, x)\alpha(x, y) d\rho(x, y) \quad s \in X && (5.5)\\ 0 &= \int_{X\times Y} K_{\mathcal{B}}(s, x)\alpha(x, y) d\rho(x, y) \quad s \in X. && (5.6) \end{aligned}$$

The proof of this theorem is given in the following subsection. A few important remarks are in order. First, the theorem gives a general quantitative version of the representer theorem. The generality is obtained by considering the population case which subsumes the sample case if the measure ρ is the empirical measure $\rho_{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$. In this case, the integral reduces to a finite sum and we recover the well known result that $f^\lambda_{\$} vz = \sum_{i=1}^n \alpha_i K_{x_i}$, where the x_i form the training set. Moreover, the solution is quantitatively characterized since the coefficients α are given by equation (5.4) involving the subgradient. For differentiable loss functions in the sample case, equation (5.4) reduces to

$$\alpha_i = \ell'(y_i, f^\lambda_{\mathbf{z}}(x_i) + g^\lambda_{\mathbf{z}}(x_i)),$$

where ℓ' denotes the derivative with respect to the second variable [Gir98, Wah98]. Second, if $\{\psi_i\}_{i=1}^m$ is a base for $\mathcal{B}$, the offset part of the solution can be written as $g^\lambda = \sum_{i=1}^m d_i\psi_i$, where the coefficients d_i are again constrained by equation (5.4). A discussion on how to solve explicitly equation (5.4) can be found in [Wah98]. Furthermore, the presence of $\mathcal{B}$ induces a system of linear constraints on the coefficients α_i expressed by equation (5.6) that, for $\mathcal{B} = \mathbb{R}$, reduces to the well known condition

$$\sum_{i=1}^n \alpha_i = 0.$$

We stress that, unlike previous works, the above equation has been derived without introducing the dual formulation. Finally, we discuss the role of assumption 3) in definition 1. From the proof, it is apparent that this assumption is needed to ensure the continuity of the first term in the Tikhonov functional which in the sample case is trivially guaranteed. Therefore, for the sample case theorem 16 holds for any convex loss function. In particular,

$L^q(Z, \rho_{\mathbf{z}}) = \mathbb{R}^n$ and the condition $\alpha \in L^q(Z, \rho_{\mathbf{z}})$ is always satisfied. Back to the population case, if $\ell(y, \cdot)$ is Lipschitz on $\mathbb{R}$ with a Lipschitz constant independent of y and

$$\int_{X \times Y} \ell(y, 0) d\rho(x, y) < +\infty,$$

one can choose $p = 1$, so that $q = +\infty$ and condition $\alpha \in L^\infty(Z, \rho)$ means that α is bounded. For the square loss, clearly $p = 2$, so that $q = 2$ and α is square-integrable. As shown by [Ste03], for classification and compact X, one can again remove assumption 3) of definition 1 using the fact that a convex function is locally Lipschitz and the range of possible y is bounded. The following corollary is the restatement of the representer theorem without offset space.

Corollary 10. *With the assumptions of theorem* 16*, let* $f^\lambda \in \mathcal{H}$ *then*

$$f^\lambda \in \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ \int_{X \times Y} \ell(y, f(x)) d\rho(x, y) + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$

if and only if there is $\alpha \in L^q(Z, \rho)$ *satisfying*

$$\begin{aligned} \alpha(x, y) &\in (\partial \ell)(y, f^\lambda(x)) \qquad (x, y) \in X \times Y \text{ a.e.} \\ f^\lambda(s) &= -\frac{1}{2\lambda} \int_{X \times Y} K(s, x) \alpha(x, y) d\rho(x, y) \quad s \in X. \end{aligned}$$

5.2.1 Proof of the main theorem

Before giving the proof of the theorem we discuss the proof structure, which aside from some technicalities is very simple, and is based on two lemmas. The Tikhonov functional $\mathcal{E}(f + g) + \lambda \|f\|_{\mathcal{H}}^2$ is a convex map on $\mathcal{H} \times \mathcal{B}$, so (f^λ, g^λ) is a minimizer of the Tikhonov functional if and only if $(0, 0)$ is in its subgradient, which is a subset of $\mathcal{H} \times \mathcal{B}$. Using linearity, the computation of the subgradient of the Tikhonov functional reduces to the computation of the subgradient of $\mathcal{E}(f + g)$ and $\|f\|_{\mathcal{H}}^2$ respectively. Since the latter functional is differentiable, the subgradient evaluation is straightforward. Some care is needed for the subgradient of the former. First, we rewrite it as an integral functional on $L^p(Z, \rho)$ and then use a fundamental result of convex analysis to interchange the integral and the subgradient.

Proof of theorem 16. Clearly, the functional $\lambda \|f\|_{\mathcal{H}}^2$ is continuous and, by lemma 5, the functional $\mathcal{E}(f + g)$ is continuous and finite. So, from item 5 of proposition 11 in the appendix, one has that

$$\partial \left(\mathcal{E}(f + g) + \lambda \|f\|_{\mathcal{H}}^2 \right) = \partial(\mathcal{E}(f + g)) + \lambda \partial(\|f\|_{\mathcal{H}}^2).$$

Now, the map

$$(f, g) \to \|f\|_{\mathcal{H}}^2$$

is differentiable with derivative $(2f, 0)$ and, therefore, by item 1 of proposition 11 in the appendix,

$$\partial(\|f\|_{\mathcal{H}}^2) = \{(2f, 0)\}. \tag{5.7}$$

The main difficulty is the evaluation of the subgradient of the map $\mathcal{E}(f + g)$ given in lemma 6. By means of this lemma we obtain that the elements of the subgradient of $\mathcal{E}(f + g)$ at (f, g) are of the form

$$\left(\int_{X\times Y} K(x, \cdot)\alpha(x, y)d\rho(x, y), \int_{X\times Y} K_{\mathcal{B}}(x, \cdot)\alpha(x, y)d\rho(x, y) \right), \tag{5.8}$$

where $\alpha \in L^q(Z, \rho)$ satisfies

$$\alpha(x, y) \in (\partial\ell)(y, f(x) + g(x)) \tag{5.9}$$

for ρ-almost all $(x, y) \in X \times Y$. Now, by combining equation (5.7) and equation (5.8), we have that the elements of the subgradient of $\mathcal{E}(f + g) + \lambda \|f\|_{\mathcal{H}}^2$ at point (f, g) are of the form

$$\left(\int_{X\times Y} K(x, \cdot)\alpha(x, y)d\rho(x, y) + 2\lambda f, \int_{X\times Y} K_{\mathcal{B}}(x, \cdot)\alpha(x, y)d\rho(x, y) \right). \tag{5.10}$$

where $\alpha \in L^q(Z, \rho)$ satisfies equation (5.9). From item 3 of proposition 11 in the appendix, we have that an element $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$ is a minimizer of $\mathcal{E}(f + g) + \lambda \|f\|_{\mathcal{H}}^2$ if and only if $(0, 0)$ belongs to the subgradient evaluated at (f^λ, g^λ). Using equation (5.10), one has that

$$f^\lambda(s) = -\frac{1}{2\lambda} \int_{X\times Y} \alpha(x, y)K(x, s)d\rho(x, y)$$

$$\int_{X\times Y} \alpha(x, y)K_{\mathcal{B}}(x, s)d\rho(x, y) = 0.$$

where, by means of equation (5.9), $\alpha \in L^q(Z, \rho)$ satisfies equation (5.4). This ends the proof. □

Before computing the subgradient of the map $\mathcal{E}(f + g)$ in lemma 6, we need to extend the definition of expected risk on $L^p(Z, \rho)$. First of all, we let

$$\mathcal{E}_0(u) = \int_{X\times Y} \ell(y, u(x, y))\, d\rho(x, y) \quad u \in L^p(Z, \rho),$$

so that $\mathcal{E}(f + g) = \mathcal{E}_0(J(f, g))$ where $J : \mathcal{H} \times \mathcal{B} \to L^p(Z, \rho)$ is the linear map

$$J(f, g) = f + g,$$

(the function $f + g$ is viewed in a natural way as a function on Z). The following lemma collects some technical facts on $\mathcal{E}_0$ and J.

Lemma 5. *With the above notations,*

1. *the functional $\mathcal{E}_0 : L^p(Z,\rho) \to [0,+\infty[$ is well-defined and continuous;*
2. *the operator $J : \mathcal{H} \times \mathcal{B} \to L^p(Z,\rho)$ is well-defined and continuous.*

Proof. Since the loss function ℓ can be regarded as function on $Z \times \mathbb{R}$, that is, $\ell(z,w) = \ell(y,w)$ where $z = (x,y)$, one has that $\mathcal{E}_0(u)$ is the Nemitski functional associated to ℓ (see Appendix), that is,

$$\mathcal{E}_0(u) = \int_Z \ell(z,u(z))\, d\rho(z) \quad u \in L^p(Z,\rho).$$

We claim that $\mathcal{E}_0(u)$ is finite. Indeed, given $u \in L^p(Z,\rho)$, by assumption 3 in definition 1,

$$\int_{X\times Y} \ell(y,u(z))\, d\rho(x,y) \;\leq\; \int_{X\times Y} a(y) + b|u(z)|^p\, d\rho(x,y) < +\infty.$$

The proof that $\mathcal{E}_0$ is continuous can be found in proposition III.5.1 of [ET83b]. In order to prove the second item, we let $f \in \mathcal{H}$. Then, by assumption 1 the kernel is bounded and

$$\begin{aligned}\int_{X\times Y} |f(x)|^p\, d\rho(x,y) &= \int_{X\times Y} |\,\langle f, K_x\rangle_{\mathcal{H}}\,|^p\, d\rho(x,y)\\ &\leq \|f\|_{\mathcal{H}}^p \int_{X\times Y} K(x,x)^{\frac{p}{2}}\, d\rho(x,y)\\ &= \kappa\,\|f\|_{\mathcal{H}}^p < +\infty.\end{aligned}$$

In particular, the function $(x,y) \mapsto f(x)$ is in $L^p(Z,\rho)$ and $\|f\|_{L^p} \leq \kappa\,\|f\|_{\mathcal{H}}$. The same relation clearly holds for $g \in \mathcal{B}$ (where we let $\kappa_{\mathcal{B}} \geq \sup_{x\in X}\sqrt{K(x,x)}$). It follows that J is well defined and

$$\|f+g\|_{L^p} \leq \kappa\,\|f\|_{\mathcal{H}} + \kappa_{\mathcal{B}}\,\|g\|_{\mathcal{B}}\,.$$

Since J is linear, it follows that J is continuous. □

Finally, the following lemma computes the subgradient of $I = \mathcal{E}_0 \circ J$.

Lemma 6. *With the above notations, let $(f,g) \in \mathcal{H}\times\mathcal{B}$, then $(\phi,\psi) \in \partial(\mathcal{E}_0 \circ J)(f,g)$ if and only if there is $\alpha \in L^q(Z,\rho)$ such that*

$$\begin{aligned}\alpha(x,y) &\in (\partial\ell)(y, f(x)+g(x)) \quad (x,y) \in X\times Y \text{ a.e.}\\ \phi(s) &= \int_{X\times Y} K(s,x)\alpha(x,y)d\rho(x,y) \quad s \in X\\ \psi(s) &= \int_{X\times Y} K_{\mathcal{B}}(s,x)\alpha(x,y)d\rho(x,y) \quad s \in X.\end{aligned}$$

Proof. Since $\mathcal{E}_0$ is finite and continuous in $0 = J(0)$, by point 6 of proposition 11 in the appendix, we know that

$$\partial(\mathcal{E}_0 \circ J)(f,g) = J^*(\partial\mathcal{E}_0)(J(f,g)), \tag{5.11}$$

where $J^* : L^q(Z,\rho) \to \mathcal{H} \times \mathcal{B}$ is the adjoint of J, that is,

$$\langle J^*\alpha, (f,g)\rangle_{\mathcal{H}\times\mathcal{B}} = \int_{X\times Y} \alpha(x,y)J(f,g)(x,y)\; d\rho(x,y).$$

First of all, we compute $\partial\mathcal{E}_0$. Since $\mathcal{E}_0(0) < +\infty$, we can apply proposition 12 so that, given $u \in L^p(Z,\rho)$, then $\alpha \in (\partial\mathcal{E}_0)(u)$ if and only if $\alpha \in L^q(Z,\rho)$ and

$$\alpha(z) \in (\partial\ell)(y, u(x,y)),$$

for ρ-almost all $(x,y) \in X \times Y$. We now compute the adjoint of J. Let $\alpha \in L^q(Z,\rho)$ and $(\phi,\psi) = J^*\alpha \in \mathcal{H} \times \mathcal{B}$. Using the reproducing property of $\mathcal{H}$ and the definition of J^* we can write

$$\begin{aligned} \phi(s) &= \langle \phi, K_s\rangle_{\mathcal{H}} \\ &= \langle J^*\alpha, (K_s,0)\rangle_{\mathcal{H}\times\mathcal{B}} = \langle \alpha, J(K_s,0)\rangle_{L^2(Z,\rho)}. \end{aligned}$$

Writing the scalar product explicitly we then find

$$\phi(s) = \int_{X\times Y} K(s,x)\alpha(x,y)d\rho(x,y).$$

Reasoning in the same way we find that

$$\psi(s) = \int_{X\times Y} K_{\mathcal{B}}(s,x)\alpha(x,y)d\rho(x,y).$$

Replacing the above formulas in equation (5.11), we have the thesis. □

5.3 Dealing with the Offset Space

In this section we deal with the offset term which often appears in regularized solutions. We first motivate our analysis, then state and discuss our main result on this issue. Finally, we give the proofs.

5.3.1 Motivations

In the previous section we minimized a Tikhonov functional on the set $\mathcal{H} \times \mathcal{B}$, dealing explicitly with the possible presence of an offset term in the form of the solution. Typical examples in which offset spaces arise are Support Vector Machine algorithms [Vap98], where the offset term is a constant accounting for the translation invariance of the separating hyperplane, and penalization methods [Wah90], where the offset space is the kernel space of the penalization operator. In this section we show that under very weak conditions the presence of an offset term is equivalent to solving a standard regularization problem with a seminorm [Wah90]. The fact that the estimator is $f^\lambda(x) + g^\lambda(x)$ (for regression) or $\operatorname{sign}\big(f^\lambda(x) + g^\lambda(x)\big)$ (for classification) suggests to replace $\mathcal{H} \times \mathcal{B}$ with the sum

$$\mathcal{S} = \mathcal{H} + \mathcal{B} = \{f + g \,|\, f \in \mathcal{H},\ g \in \mathcal{B}\}.$$

The hypothesis space $\mathcal{S}$ is a space of functions on X and, in particular, a RKH space, the kernel being the sum of the kernels of $\mathcal{H}$ and $\mathcal{B}$. In this section we show that the minimization of a Tikhonov functional on $\mathcal{H}\times\mathcal{B}$ is essentially equivalent to the minimization of an appropriate functional on $\mathcal{S}$. This provides a rigorous derivation of the following facts.

1. The equivalent functional on $\mathcal{S}$ is also a Tikhonov functional. The penalty term is a seminorm penalizing the functions in $\mathcal{S}$ orthogonal to $\mathcal{B}$ only.
2. The estimator given by the minimization of the Tikhonov functional on $\mathcal{S}$ depends only on the kernel sum.

Finally, we notice that the norm of $\mathcal{B}$ (hence the kernel $K_\mathcal{B}$) plays no role in the functional

$$\mathcal{E}(f + g) + \lambda \|f\|^2_\mathcal{H},$$

that is, all kernels, whose corresponding RKH space is $\mathcal{B}$ as a vector space, give rise to the same minimizers (f^λ, g^λ). This fact is confirmed by equation (5.14) below (see also equation (5.16)).

5.3.2 Main theorem

We recall that the norm in $\mathcal{S}$ is given by

$$\|f + g\|^2_\mathcal{S} = \inf_{\substack{f' \in \mathcal{H}, g' \in \mathcal{B} \\ f+g=f'+g'}} \Big(\|f'\|^2_\mathcal{H} + \|g'\|^2_\mathcal{B} \Big) \tag{5.12}$$

and, with respect to this norm, $\mathcal{S}$ is a RKH space on X with kernel $K + K_\mathcal{B}$ [Sch64]. We are now ready to state the following result.

Theorem 17. *Let Q be the orthogonal projection on the closed subspace of $\mathcal{S}$*

$$\mathcal{S}_0 = \{s \in \mathcal{S} \mid \langle s, g\rangle_{\mathcal{S}} = 0 \quad \forall g \in \mathcal{B}\},$$

that is the subset of functions orthogonal to $\mathcal{B}$ w.r.t. the scalar product in $\mathcal{S}$. We have the following facts.

1. *If $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$ is a solution of the problem*

$$\min_{(f,g)\in\mathcal{H}\times\mathcal{B}} \{\mathcal{E}(f+g) + \lambda \|f\|_{\mathcal{H}}^2\},$$

then $s^\lambda = f^\lambda + g^\lambda \in \mathcal{S}$ is a solution of the problem

$$\min_{s\in\mathcal{S}} \{\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2\}$$

and $f^\lambda = Qs^\lambda$.

2. *If $s^\lambda \in \mathcal{S}$ is a solution of the problem*

$$\min_{s\in\mathcal{S}} \{\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2\},$$

let $f^\lambda = Qs^\lambda$ and $g^\lambda = s^\lambda - Qs^\lambda$, then

$$\mathcal{E}(f^\lambda + g^\lambda) + \lambda \left\|f^\lambda\right\|_{\mathcal{H}}^2 = \inf_{(f,g)\in\mathcal{H}\times\mathcal{B}} \{\mathcal{E}(f+g) + \lambda \|f\|_{\mathcal{H}}^2\}.$$

In particular, if $g^\lambda \in \mathcal{B}$, then $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$ is a minimizer of $\mathcal{E}(f+g) + \lambda \|f\|_{\mathcal{H}}^2$.

Before giving the proof in the following subsection we comment on this result. First, notice that if $\mathcal{H} \cap \mathcal{B} = \{0\}$ then $\mathcal{S} = \mathcal{H} \times \mathcal{B}$ and

$$\|f+g\|_{\mathcal{S}}^2 = \|f\|_{\mathcal{H}}^2 + \|g\|_{\mathcal{B}}^2 .$$

In this case the theorem is trivial. However, in the arbitrary case care is needed because there are functions in $\mathcal{H}$ not orthogonal to $\mathcal{B}$. Moreover, the norm $\|\cdot\|_{\mathcal{S}}$ restricted to $\mathcal{H}$ and $\mathcal{B}$ could be different from $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{B}}$; in particular, it could happen that $(\mathcal{B}^\perp)^\perp \neq \mathcal{B}$, where the orthogonality $^\perp$ is meant with respect to the inner product in $\mathcal{S}$. This pathology is at the root of the fact that there are cases in which the problem

$$\min_{s\in\mathcal{S}} \{\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2\}$$

has a solution, whereas the functional $\mathcal{E}(f+g) + \lambda \|f\|_{\mathcal{H}}^2$ does not admit a minimizer on $\mathcal{H} \times \mathcal{B}$ (see example below). In practice, since $\mathcal{H} \cap \mathcal{B}$ in most applications is finite

dimensional, this pathology does not occur and the minimization problem on $\mathcal{H} \times \mathcal{B}$ is fully equivalent to the one on $\mathcal{S}$. Second, the advantage of using the penalty term $\|f\|^2_{\mathcal{H}}$ instead of $\|Qs\|^2_{\mathcal{S}}$ is that one can solve the minimization problem without knowing the explicit form of the projection Q. Conversely, the space $\mathcal{S}$ is the natural space to address theoretical issues. Third, we observe that since the proof does not depend on the convexity of the loss function, the theorem holds for arbitrary (positive) loss functions. However, if ℓ satisfies the hypotheses of definition 1, from theorem 16 it follows that the minimizer s^λ of $\mathcal{E}(s) + \lambda \|Qs\|^2_{\mathcal{S}}$ is of the form

$$s^\lambda(s) = \int_{X\times Y} \alpha(x,y)\,(K(x,s) + K_{\mathcal{B}}(x,s))\; d\rho(x,y) + g^\lambda(s) \tag{5.13}$$

$$= \int_{X\times Y} \alpha(x,y) K(x,s)\, d\rho(x,y) + g^\lambda(s) \tag{5.14}$$

where $g^\lambda \in \overline{\mathcal{B}}$ and $\alpha \in L^q(Z,\rho)$ satisfies

$$\alpha(x,y) \in (\partial\ell)(y, s^\lambda(x)) \tag{5.15}$$

$$\int_{X\times Y} \alpha(x,y) K_{\mathcal{B}}(x,s) = 0. \tag{5.16}$$

In particular, this implies that, given $h \in \mathcal{B}$, one can replace the kernel K with $K(x,s) + h(x)h(s)$, without changing the form of the minimizer s^λ. For example if $\mathcal{B}$ is the set of constant functions, the two kernels $K(x,s) = x \cdot s$ and $K(x,s) = x \cdot s + 1$ are equivalent since both penalize the functions orthogonal to 1, that is the space of linear functions.

5.3.3 Proof

Before giving the proof of theorem 17 we need to prove the following technical lemma. To this purpose we recall that $\mathcal{S}_0$ was defined as

$$\mathcal{S}_0 = \{s \in \mathcal{S} \mid \langle s, g\rangle_{\mathcal{S}} = 0 \;\; \forall g \in \mathcal{B}\},$$

and Q was the corresponding orthogonal projection from $\mathcal{S}$ onto $\mathcal{S}_0$. Moreover we let $\mathcal{H}_0$ be the closed subspace of $\mathcal{H}$ given by

$$\mathcal{H}_0 = \{f \in \mathcal{H} \mid \langle f, h\rangle_{\mathcal{H}} = 0 \;\; \forall h \in \mathcal{H} \cap \mathcal{B}\}$$

and P be the corresponding orthogonal projection from $\mathcal{H}$ onto $\mathcal{H}_0$. In order to prove the main theorem we need the following technical lemma that characterizes the space $\mathcal{S}_0$.

Lemma 7. *Let $s = f + g \in \mathcal{S}$ with $f \in \mathcal{H}$ and $g \in \mathcal{B}$, then*

$$Qs = Pf \tag{5.17}$$

$$\|Qs\|_{\mathcal{S}} = \|Pf\|_{\mathcal{H}} \tag{5.18}$$

and there is a sequence $(f_n, g_n) \in \mathcal{H} \times \mathcal{B}$ *such that*

$$\lim_{n\to\infty} \|Pf - f_n\|_{\mathcal{H}} \;=\; 0 \tag{5.19}$$

with $f_n + g_n = s$.

Equations (5.17) and (5.18) show that $\mathcal{S}_0$ and $\mathcal{H}_0$ are the same Hilbert space and, in particular, $Qs \in \mathcal{H}$. However, in general, it could happen that $s - Qs \notin \mathcal{B}$. Equation (5.19) is a technical trick to overcome this pathology.

Proof of lemma 7. To give the proof of the lemma we need some preliminary facts. Let $\mathcal{K}$ be the closed subspace of $\mathcal{H} \times \mathcal{B}$

$$\mathcal{K} = \{(f, g) \in \mathcal{H} \times \mathcal{B} \,|\, (f, h)_{\mathcal{H}} = (g, h)_{\mathcal{B}} \;\forall h \in \mathcal{H} \cap \mathcal{B}\}.$$

It is known [Sch64] that, given $s \in \mathcal{S}$, there is a unique $(f, g) \in \mathcal{K}$ such that $s = f + g$. Moreover for all $(f', g') \in \mathcal{H} \times \mathcal{B}$,

$$\langle s, f' + g' \rangle_{\mathcal{S}} = \langle f, f' \rangle_{\mathcal{H}} + \langle g, g' \rangle_{\mathcal{B}}\,. \tag{5.20}$$

From equation (5.12) one has that

$$\|f\|_{\mathcal{S}} \leq \|f\|_{\mathcal{H}} \quad f \in \mathcal{H} \tag{5.21}$$

First of all we claim that $\mathcal{H}_0 \subset \mathcal{S}_0$. Clearly, if $f \in \mathcal{H}_0$, then $(f, 0) \in \mathcal{K}$ and, by equation (5.20), for all $g' \in \mathcal{B}$,

$$\langle f + 0, 0 + g' \rangle_{\mathcal{S}} = \langle f, 0 \rangle_{\mathcal{H}} + \langle 0, g' \rangle_{\mathcal{B}} = 0,$$

that is $f \in \mathcal{S}_0$. This shows the claim. Moreover,

$$\|f\|^2_{\mathcal{S}} = \langle f + 0, f + 0 \rangle_{\mathcal{S}} = \langle f, f \rangle_{\mathcal{H}} = \|f\|^2_{\mathcal{H}}\,. \tag{5.22}$$

Let $s = f + g$ with $f \in \mathcal{H}$ and $g \in \mathcal{B}$. Clearly, $f = Pf + h$ where $h \in \mathcal{H}_0^{\perp} = ((\mathcal{H} \cap \mathcal{B})^{\perp})^{\perp} = \mathcal{H} \bar{\cap} \mathcal{B}$ (here $^{\perp}$ denotes the orthogonal complement with respect to the scalar product of $\mathcal{H}$). It follows that there is a sequence $h_n \in \mathcal{H} \cap \mathcal{B}$ such that

$$\lim_{n\to\infty} \|h - h_n\|_{\mathcal{H}} = 0. \tag{5.23}$$

Since, by equation (5.21), $\|h - h_n\|_{\mathcal{S}} \leq \|h - h_n\|_{\mathcal{H}}$ and Q is continuous, it follows that $Qh = \lim_{n\to\infty} Qh_n = 0$, since $Qh_n = 0$. The statements of the theorem easily follow from the above facts. Indeed

$$Qs = Q(Pf + h + g) = QPf = Pf,$$

since $Pf \in \mathcal{H}_0 \subset \mathcal{S}_0$, and equation (5.17) is proved. Equation (5.18) follows from equation (5.22). Finally let now $f_n = Pf + h - h_n$ and $g_n = g + h_n$. Clearly, $f_n + g_n = f + g = s$, $f_n \in \mathcal{H}$ and $g_n \in \mathcal{B}$ and moreover equation (5.19) follows from equation (5.23). □

We are now ready to prove the main theorem of this section.

Proof of theorem 17. First of all we note the following facts. Let $f \in \mathcal{H}$, $g \in \mathcal{B}$ and $s = f + g \in \mathcal{S}$. By equation (5.18)

$$\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2 = \mathcal{E}(f+g) + \lambda \|Pf\|_{\mathcal{H}}^2 \tag{5.24}$$

Let $(f_n, g_n) \in \mathcal{H} \times \mathcal{B}$ as in lemma 7, then

$$\mathcal{E}(f+g) + \lambda \|Pf\|_{\mathcal{H}}^2 = \lim_n \left(\mathcal{E}(f_n + g_n) + \lambda \|f_n\|_{\mathcal{H}}^2\right).$$

From the above equalities it follows that

$$\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2 = \lim_n \left(\mathcal{E}(f_n + g_n) + \lambda \|f_n\|_{\mathcal{H}}^2\right). \tag{5.25}$$

We can now prove the first part of the theorem. Assume that $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$ is a minimizer of $\mathcal{E}(f+g) + \lambda \|f\|_{\mathcal{H}}^2$ and let $s^\lambda = f^\lambda + g^\lambda$. From equation (5.25) and the definition of minimizer, one has that, for all $s \in \mathcal{S}$,

$$\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2 \geq \mathcal{E}(f^\lambda + g^\lambda) + \lambda \left\|f^\lambda\right\|_{\mathcal{H}}^2. \tag{5.26}$$

In particular with the choice $s = s^\lambda$, by means of equation (5.18), one has that

$$\|Qs\|_{\mathcal{S}} = \left\|Pf^\lambda\right\|_{\mathcal{H}} \geq \left\|f^\lambda\right\|_{\mathcal{H}},$$

and, hence, that $Qs^\lambda = Pf^\lambda = f^\lambda$. Therefore, it follows that

$$\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2 \geq \mathcal{E}(s^\lambda) + \lambda \left\|Qs^\lambda\right\|_{\mathcal{S}}^2,$$

that is, s^λ is a minimizer of $\mathcal{E}(s) + \lambda \|Ps\|_{\mathcal{S}}^2$. Before proving the second part of the theorem we note that the following inequality follows as a simple consequence of the definition of projection.

$$\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2 = \mathcal{E}(f+g) + \lambda \|Pf\|_{\mathcal{H}}^2 \leq \mathcal{E}(f+g) + \lambda \|f\|_{\mathcal{H}}^2. \tag{5.27}$$

Assume now that $s^\lambda \in \mathcal{S}$ is a minimizer of $\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2$. Let $f^\lambda = Qs^\lambda$ and $g^\lambda = s - f^\lambda$, then, by equation (5.27) and equation (5.18), it follows that

$$\mathcal{E}(f^\lambda + g^\lambda) + \lambda \left\|f^\lambda\right\|_{\mathcal{H}}^2 \leq \inf_{(f,g) \in \mathcal{H} \times \mathcal{B}} \{\mathcal{E}(f+g) + \lambda \|f\|_{\mathcal{H}}^2\}.$$

However, using equation (5.25) with $s = f^\lambda + g^\lambda$, one has that

$$\mathcal{E}(f^\lambda + g^\lambda) + \lambda \left\|f^\lambda\right\|_{\mathcal{H}}^2 \geq \inf_{(f,g) \in \mathcal{H} \times \mathcal{B}} \{\mathcal{E}(f+g) + \lambda \|f\|_{\mathcal{H}}^2\}.$$

So $\mathcal{E}(f^\lambda + g^\lambda) + \lambda \left\|f^\lambda\right\|_{\mathcal{H}}^2$ is the infimum of $\mathcal{E}(f+g) + \lambda \|f\|_{\mathcal{H}}^2$ on $\mathcal{H} \times \mathcal{B}$. Clearly, if $g^\lambda \in \mathcal{B}$, it follows that (f^λ, g^λ) is a minimizer of $\mathcal{E}(f+g) + \lambda \|f\|_{\mathcal{H}}^2$. □

5.3.4 A counterexample

The following example shows that in some pathological framework the minimization on $\mathcal{H} \times \mathcal{B}$ is not equivalent to the one on $\mathcal{S} = \mathcal{H} + \mathcal{B}$.

Example 5. *Let $\mathcal{H} = \ell_2 = \{f = (f_n)_{n\in\mathbb{N}} \,|\, \sum_n f_n^2 < +\infty\}$. The space n_2 is a RKH space on $\mathbb{N}$ with respect to the kernel $K(n,m) = \delta_{n,m}$. Let $\mathcal{B} = \{f \in \ell_2 \,|\, \sum_n n^2 f_n^2 < +\infty\}$ with the scalar product*

$$\langle f, g\rangle_{\mathcal{B}} = \sum_n n^2 f_n g_n.$$

The space $\mathcal{B}$ is a RKH space with respect to the kernel $K_{\mathcal{B}}(n,m) = \frac{1}{n^2}\delta_{n,m}$. Clearly, $\mathcal{B} \subset \mathcal{H}$, so that $\mathcal{H}\cap\mathcal{B} = \mathcal{B}$, which is not closed in $\mathcal{H}$. Since $\mathcal{B}$ is dense in $\mathcal{H}$, $P = 0$ and, by lemma 7, $Q = 0$. Let $\ell(y, f(x))$ be the square loss function and choose $h = (h_n)_{n\in\mathbb{N}} \in \mathcal{H}$ such that $h \notin \mathcal{B}$. Let $\rho(n, y) = \delta(y - h_n)$ so that

$$\mathcal{E}(s) = \|s - h\|_{\mathcal{S}}^2,$$

then

$$\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2 = \|s - h\|_{\mathcal{S}}^2,$$

and the minimizer is $s^\lambda = h$. Moreover, by our theorem, one has that

$$\inf_{f\in\mathcal{H}, g\in\mathcal{B}} \{\mathcal{E}(f + g) + \lambda \|f\|_{\mathcal{H}}^2\} = \mathcal{E}(s^\lambda) + \lambda \left\|Qs^\lambda\right\|_{\mathcal{S}}^2 = 0.$$

If $(f^\lambda, g^\lambda) \in \mathcal{H}\times\mathcal{B}$ were a minimizer, then $f^\lambda = 0$ and, hence, $g^\lambda = h$, but this is impossible since $h \notin \mathcal{B}$.

5.4 Existence and uniqueness

We now discuss existence and uniqueness of the regularized solution in $\mathcal{S}$. Before stating and proving the main results we summarize our findings and show that if the offset space is empty both existence and uniqueness are easily obtained. Our analysis extends existence to all cases of interest under some weak assumptions on the kernel and the loss function for both regression and classification. Uniqueness depends critically on the convexity assumption. For strictly convex functions we prove that the solution is unique if and only if the offset space satisfies suitable conditions, fulfilled in the case of constant offsets. For loss functions which are not strictly convex we limit our attention to the hinge loss and show that the solution is unique unless some particular conditions on the number and location of the support vectors are met. In [BC00, BC03] similar results were obtained considering the dual formulation of the minimization problem. If the offset space is empty, strict convexity and coerciveness of the penalty term trivially imply both existence and uniqueness. Indeed, we have the following proposition.

Proposition 5. *Given $\lambda > 0$, there exists a unique solution of the problem*

$$\min_{f \in \mathcal{H}} \left(\mathcal{E}(f) + \lambda \|f\|_{\mathcal{H}}^2 \right).$$

Proof. The function $\left(\mathcal{E}(f) + \lambda \|f\|_{\mathcal{H}}^2 \right)$ is strictly convex and continuous. Moreover

$$\mathcal{E}(f) + \lambda \|f\|_{\mathcal{H}}^2 \geq \lambda \|f\|_{\mathcal{H}}^2 \to +\infty$$

if $\|f\|_{\mathcal{H}}$ goes to $+\infty$. From item 4 of proposition 11 in the appendix both existence and uniqueness follow. □

5.4.1 Existence

We now consider existence. If $\mathcal{B}$ is not trivial, there are no general results (see [Wah90] for a discussion on this subject). However, if $\mathcal{B}$ is the set of constant functions, we derive existence of the solution in two different settings. The first proposition holds only for classification under the assumption that the loss function ℓ goes to infinity when $yf(x)$ goes to $-\infty$ (see Condition 1 of proposition 6 below). Similar results were obtained in [Ste04]. We let ρ_X be the marginal measure on X associated to ρ and $\operatorname{supp} \rho_X$ its support.

Proposition 6. *Let the kernel K satisfy assumption 1. Moreover assume that the following conditions hold*

1. $\lim_{w \to -\infty} \ell(1, w) = +\infty$ *and* $\lim_{w \to +\infty} \ell(-1, w) = +\infty$
2. $\rho(X \times \{1\}) > 0$ *and* $\rho(X \times \{-1\}) > 0$

Then there is at least one solution of the problem

$$\min_{s \in \mathcal{S}} \left(\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2 \right),$$

where $\mathcal{S} = \mathcal{H} + \mathbb{R}$.

We observe that item *2.* has a very natural interpretation in the discrete setting where it simply amounts to have one example for each class. This condition is needed since item *1.* does not requires that ℓ goes to $+\infty$ when $yf(x)$ goes to $+\infty$. A typical example of loss function satisfying item *1.* is the hinge loss. The second result holds both for regression and classification, but it requires the loss function going to infinity when $f(x)$ goes to $\pm\infty$, uniformly in y (compare assumption *1.* of proposition 7 and assumption *1.* of proposition 6).

Proposition 7. *Let the kernel K satisfy assumption 1. Moreover assume that the following conditions hold*

$$\lim_{w\to\pm\infty}(\inf_{y\in Y}\ell(y,w))=+\infty.$$

Then there is at least one solution of the problem

$$\min_{s\in\mathcal{S}}\left(\mathcal{E}(s)+\lambda\left\|Qs\right\|_{\mathcal{S}}^{2}\right),$$

where $\mathcal{S}=\mathcal{H}+\mathbb{R}$.

We observe that for classification with symmetric loss functions, as the square loss function, this proposition gives a sharper result than proposition 6. We now prove proposition 6 and omit the proof of proposition 7 since it is essentially the same.

Proof of proposition 6. The idea of the proof is to show that the functional we have to minimize goes to $+\infty$ when $\left\|s\right\|_{\mathcal{S}}$ goes to $+\infty$. To this aim, let

$$\alpha=\min\{\rho\left(X\times\{1\}\right),\rho\left(X\times\{-1\}\right)\}.$$

By assumption *3*, $\alpha>0$. For a fixed $M>0$, we are looking for $R>0$ such that for all $s\in\mathcal{S}$ with $\left\|s\right\|_{\mathcal{S}}\geq R$,

$$\mathcal{E}(s)+\lambda\left\|Qs\right\|_{\mathcal{S}}^{2}\geq M.$$

Due to assumption *1*, there is $r>0$ such that, for all $w\leq -r$, $\ell(1,w)\geq\frac{M}{\alpha}$ and, for all $w\geq r$, $\ell(-1,w)\geq\frac{M}{\alpha}$. We now let $R=\max\{2(1+\kappa)\sqrt{\frac{M}{\lambda}},2r\}$ and choose $s\in\mathcal{S}$ with $\left\|s\right\|_{\mathcal{S}}\geq R$. If $\left\|Qs\right\|_{\mathcal{S}}=\left\|Qs\right\|_{\mathcal{H}}\geq\frac{R}{2(1+\kappa)}$, then

$$\begin{aligned}\mathcal{E}(s)+\lambda\left\|Qs\right\|_{\mathcal{S}}^{2} &\geq \lambda\left\|Qs\right\|_{\mathcal{S}}^{2}\\ &\geq \lambda(\frac{R}{2(1+\kappa)})^{2}\\ &\geq M,\end{aligned}$$

since $R\geq 2(1+\kappa)\sqrt{\frac{M}{\lambda}}$. If $\left\|Qs\right\|\leq\frac{R}{2(1+\kappa)}$, let $b=s-Qs\in\mathbb{R}$, then

$$\begin{aligned}|b| &= \left\|s-Qs\right\|_{\mathcal{S}}\\ &\geq \left\|s\right\|_{\mathcal{S}}-\left\|Qs\right\|_{\mathcal{S}}\\ &\geq R-\frac{R}{2(1+\kappa)}=R\frac{2\kappa+1}{2\kappa+2}\end{aligned}$$

Assume, for example, that $b > 0$. For all $x \in \operatorname{supp} \rho_X$

$$\begin{aligned} s(x) &= \langle Qs, K_x \rangle_{\mathcal{H}} + b \\ &\geq b - \|Qs\|_{\mathcal{H}} \|K_x\|_{\mathcal{H}} \\ &\geq R\frac{2\kappa+1}{2\kappa+2} - \frac{R}{2(1+\kappa)}\kappa \\ &\geq R\frac{\kappa+1}{2\kappa+2} \\ &= \frac{R}{2} \geq r, \end{aligned}$$

since $R \geq \frac{r}{2}$. By definition of r, one has that for all $x \in \operatorname{supp} \rho_X$

$$\ell(-1, s(x)) \geq \frac{M}{\alpha}.$$

Integrating both sides, we find

$$\int_{X\times\{-1\}} \ell(-1, s(x))\, d\rho(x, -1) \geq \frac{M}{\alpha}\rho(X \times \{-1\}) \geq M$$

from which it follows that

$$\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2 \geq M.$$

The same proof holds when $b < 0$ replacing the integration on $X\times\{-1\}$ with the integration on $X \times \{1\}$. Since M is arbitrary, we have that

$$\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2 \geq \lambda \|Qs\|_{\mathcal{S}}^2 \to +\infty.$$

Since the functional is continuous, from item 4 of proposition 11 in the appendix the existence of the minimizer follows. □

5.4.2 Uniqueness

The following proposition completely characterizes uniqueness for strictly convex functions.

Proposition 8. *Let s^λ be a solution of the problem*

$$\min_{s\in\mathcal{S}} \left(\mathcal{E}(s) + \lambda \|Qs\|_{\mathcal{S}}^2\right).$$

1. *If s' is another solution, then $Qs' = Qs^\lambda$.*
2. *If $\ell(y, \cdot)$ is strictly convex for all $y \in Y$ then all the minimizers are of the form $s^\lambda + g$, with $g \in \mathcal{S}$ such that $Qg = 0$ and $g(x) = 0$ for ρ_X-almost all $x \in X$.*

Let us comment on this proposition before providing the proof. We recall that a solution s^λ is the sum of two terms: $f^\lambda = Qs^\lambda$ which is orthogonal to $\mathcal{B}$ and $g^\lambda = s^\lambda - f^\lambda$. The uniqueness of f^λ (item *1*) is due to the strict convexity of the penalty term. Item *2* states the general conditions that should be satisfied by offset functions to obtain uniqueness on s^λ: in the sample case one has uniqueness if and only if the condition $g(x_i) = 0$ for all i implies that g is equal to zero. Clearly, if $\mathcal{B}$ is the space of constant functions uniqueness is ensured. We now give the proof of the proposition.

Proof of proposition 8. 1. Let s' another minimizer and assume that $Qs^\lambda \neq Qs'$. Then, by the strict convexity of $\|\cdot\|_{\mathcal{S}}^2$, one has that, for all $t \in]0,1[$,

$$\left\|(1-t)Qs^\lambda + tQs'\right\|_{\mathcal{S}}^2 < (1-t)\left\|Qs^\lambda\right\|_{\mathcal{S}}^2 + t\left\|Qs'\right\|_{\mathcal{S}}^2.$$

Since $\mathcal{E}(s)$ is convex, one has that

$$\mathcal{E}((1-t)s^\lambda + ts') \leq (1-t)\mathcal{E}(s^\lambda) + t\mathcal{E}(s').$$

From the above two inequalities we find

$$\begin{aligned}\mathcal{E}((1-t)s^\lambda + ts') &+ \lambda\left\|Q\left((1-t)s^\lambda + ts'\right)\right\|_{\mathcal{S}}^2 \\ &< (1-t)\left(\mathcal{E}(s^\lambda) + \lambda\left\|Qs^\lambda\right\|_{\mathcal{S}}^2\right) + t\left(\mathcal{E}(s') + \lambda\left\|Qs'\right\|_{\mathcal{S}}^2\right) \\ &= \min_{s\in\mathcal{S}}\left(\mathcal{E}(s) + \lambda\left\|Qs\right\|_{\mathcal{S}}^2\right).\end{aligned}$$

Since this is impossible, it follows that $Qs^\lambda = Qs'$.

2. Let $s' = s^\lambda + g$ with g as in item *1*. By straightforward computation we have that s' is a minimizer. It is left to show that the minimizers are only the functions written in the above form. From item *1* we have that $Qg = 0$. Let U be the measurable set

$$U = \{x \in X \mid g(x) \neq 0\} = \{x \in X \mid s'(x) \neq s^\lambda(x)\}.$$

By contradiction, let us assume that $\rho_X(U) > 0$ and, hence, $\rho(U \times Y) > 0$. Fix $t \in]0,1[$. since $\ell(y,\cdot)$ is strictly convex, for all $(x,y) \in U \times Y$, one has that

$$\ell(y,(1-t)s^\lambda(x) + ts'(x)) < (1-t)\ell(y,s^\lambda(x)) + t\ell(y,s'(x)).$$

Therefore, by integration,

$$\int_{U\times Y} \ell(y,(1-t)s^\lambda(x) + ts'(x))\,d\rho(x,y) <$$

$$< (1-t)\int_{U\times Y} \ell(y,s^\lambda(x))\,d\rho(x,y) + t\int_{U\times Y} \ell(y,s'(x))\,d\rho(x,y).$$

On the complement of $U \times Y$, we have $\ell(y, s^\lambda(x)) = \ell(y, s'(x))$, so that

$$\mathcal{E}((1-t)s^\lambda + ts') < (1-t)\mathcal{E}(s^\lambda) + t\mathcal{E}(s').$$

By the same line of reasoning of item *1*, one finds a contradiction. It follows that $\rho_X(U) = 0$, that is, $g(x) = 0$ for ρ_X-almost all $x \in X$.

□

Two important examples of convex loss functions which are not strictly convex are the hinge and the ϵ-insensitive loss. The next proposition deals with the hinge loss though a similar result can be also derived for the ϵ-insensitive loss, see [BC00]. For the sake of simplicity we develop our result in the sample case for the case of constant offset functions. In this case uniqueness of the solution is expressed as a condition on the number of support vectors of the two classes. Similar but a little bit more involved conditions can be found considering the population case.

Proposition 9. *Let $Y = \{\pm 1\}$, $\ell(y, w) = |1 - yw|_+$ and $\mathcal{B} = \mathbb{R}$. Let s^λ be a solution of*

$$\min_{s \in \mathcal{S}} \left(\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, s(x_i)) + \lambda \left\| Qs \right\|_{\mathcal{S}}^2 \right),$$

and define

$$\begin{aligned} I_+ &= \{i \,|\, y_i = 1,\ s^\lambda(x) < 1\} \qquad I_- = \{i \,|\, y_i = -1,\ s^\lambda(x) > -1\} \\ B_+ &= \{i \,|\, y_i = 1,\ s^\lambda(x_1) = 1\} \qquad B_- = \{i \,|\, y_i = -1,\ s^\lambda(x_1) = -1\}. \end{aligned}$$

The solution is unique if and only if

$$\#I_+ \neq \#I_- + \#B_- \tag{5.28}$$

and

$$\#I_- \neq \#I_+ + \#B_+, \tag{5.29}$$

where # denotes set cardinality.

Proof. Assume that s' is another solution. From item *1* of proposition 8, we have that $Qs^\lambda = Qs'$ and $s' = s^\lambda + b$. Since both functions are minimizers, one concludes that

$$\sum_{i=1}^{n} |1 - y_i s^\lambda(x_i)|_+ = \sum_{i=1}^{n} |1 - y_i s'(x_i)|_+ \tag{5.30}$$

We notice that if $yw_1 < 1$ and $yw_2 > 1$, then

$$\ell(y, (1-t)w_1 + tw_2) < (1-t)\ell(y, w_1) + t\ell(y, w_2).$$

Reasoning as in the proof of the previous proposition, one has that, for all $i \in I_+ \cup I_-$,

$$y_i s'(x_i) \leq 1$$

and, for all $i \notin (I_+ \cup I_- \cup B_+ \cup B_-)$

$$y_i s'(x_i) \geq 1.$$

Using the above two equations, it follows that equality (5.30) becomes

$$\sum_{i \in I_+ \cup I_-} (1 - y_i s^\lambda(x_i)) = \sum_{i \in I_+ \cup I_-} (1 - y_i s'(x_i)) + \sum_{i \in B_+ \cup B_-} |-by_i|_+,$$

(if the index set is empty, we let the corresponding sum be equal to 0). The above equation is equivalent to

$$\sum_{i \in I_+ \cup I_-} by_i = \sum_{i \in B_+ \cup B_-} |-by_i|_+,$$

that has a not trivial solution if and only if both the following conditions are true

1. if $b > 0$, then $\sum_{i \in I_+ \cup I_-} y_i = -\sum_{B_-} y_i$ (that is, equation (5.28) holds).
2. if $b < 0$, then $\sum_{i \in I_+ \cup I_-} y_i = \sum_{B_+} y_i$ (that is, equation (5.29) holds).

Now, if neither equation (5.28) nor equation (5.29) holds, then $b = 0$ and s^λ is unique. Conversely, assume for example that equation (5.28) holds. It is simple to check that there is $b > 0$ such that for all $i \in I_+ \cup I_-$,

$$y_i(s^\lambda(x_i) + b) \leq 1$$

and, for all $i \notin (I_+ \cup I_- \cup B_+ \cup B_-)$

$$y_i(s^\lambda(x_i) + b) \geq 1.$$

Finally, by direct computation one has that

$$\mathcal{E}(s^\lambda) = \mathcal{E}(s^\lambda + b).$$

□

If the solution is not unique, the solution family is parameterized as $s^\lambda + b$, where b runs in a closed, not necessarily bounded interval. However, if there is at least one example for each class, b lies in the bounded interval $[b_-, b_+]$ and one can easily show that

1. for the solution with $b = b_-$, equation (5.28) holds;
2. for the solution with $b = b_+$, equation (5.29) holds;
3. for the solution with $b_- < b < b_+$, both Eqs. (5.28) and (5.29) hold, from which it follows that $\#I_+ = \#I_-$ and $\#B_+ = \#B_- = 0$.

5.5 Sample Case and Support Vector Machines

We now specialize the previous results to the case in which the probability measure is the empirical distribution $\rho_{\mathbf{z}}$ and $\mathcal{B}$ is the space of constant functions ($K_{\mathcal{B}} = 1$) and discuss in detail Support Vector Machines for classification. We start recalling that, from item 2 of proposition 11 in the appendix it follows that the left and right derivatives of $\ell(y, \cdot)$ always exist and

$$(\partial\ell)(y, w) = [\ell'_-(y, w), \ell'_+(y, w)].$$

Corollary 11. *Let $\mathcal{S} = \mathcal{H} + \mathbb{R}$ and Q the projection on*

$$\{s \in \mathcal{S} \mid \langle s, 1\rangle_{\mathcal{S}} = 0\}.$$

Given $\lambda > 0$, let $f^\lambda \in \mathcal{H}$ and $b^\lambda \in \mathbb{R}$ and define $s^\lambda = f^\lambda + b^\lambda \in \mathcal{S}$, then

$$(f^\lambda, b^\lambda) \in \operatorname*{argmin}_{f\in\mathcal{H}, b\in\mathbb{R}} \left\{ \frac{1}{n}\sum_i \ell(y_i, f(x_i) + b) + \lambda \|f\|^2_{\mathcal{H}} \right\}$$

if and only if

$$\begin{aligned} s^\lambda &\in \operatorname*{argmin}_{s\in\mathcal{S}} \left\{ \frac{1}{n}\sum_i \ell(y_i, s(x_i)) + \lambda \|Qs\|^2_{\mathcal{H}} \right\} \\ f^\lambda &= Qs^\lambda. \end{aligned}$$

Moreover there are $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that

$$\begin{aligned} f^\lambda &= \sum_{i=1}^n \alpha_i K_{x_i} = \sum_{i=1}^n \alpha_i (K_{x_i} + 1) \\ \frac{-1}{2\lambda n}\ell'_+(y_i, f^\lambda(x_i) + b^\lambda) \le \alpha_i &\le \frac{-1}{2\lambda n}\ell'_-(y_i, f^\lambda(x_i) + b^\lambda) \\ \sum_{i=1}^n \alpha_i &= 0 \end{aligned}$$

We notice two facts. First, α_i can be zero only if $0 \in (\partial\ell)(y_i, f^\lambda(x_i) + b^\lambda)$ – that is, only if $f^\lambda(x_i) + b^\lambda$ is a minimizer of $\ell(y_i, \cdot)$. Therefore, *a necessary condition for obtaining sparsity is a plateau in the loss function.* A quantitative discussion on this topic can be found in [Ste03]. Second if ℓ_- and ℓ_+ are bounded by a constant $M > 0$, one has that $|\alpha_i| \le 2\lambda n M$ – that is, a sufficient conditions for box constraints on the coefficients. In the rest of this section we consider Support Vector Machines for classification showing that through

our analysis the solution is completely characterized in the primal formulation. A simple calculation for the hinge loss shows that

$$[\ell'_-(y,w), \ell'_+(y,w)] = \begin{cases} -y & \text{for } yw < 1 \\ [\min\{-y,0\}, \max\{0,-y\}] & \text{for } yw = 1 \\ 0 & \text{for } yw > 1 \end{cases} . \tag{5.31}$$

To be consistent with the notation used in the literature, we let $C = \frac{1}{2\lambda n}$ and factorize the labels y_i from the coefficients α_i. Then, according to the above corollary, the solution of the SVM algorithm is given by

$$s^\lambda = \sum_{i=1}^{n} \alpha_i y_i K_{x_i} + b^\lambda$$

where the set $(\alpha_1, \ldots, \alpha_n, b^\lambda)$ solves the following algebraic system of inequalities

$$\begin{aligned} 0 \le \alpha_i \le C \quad &\text{if} \quad y_i \left(\sum_{j=1}^{n} \alpha_j y_j K(x_i, x_j) + b^\lambda \right) = 1 \\ \alpha_i = 0 \quad &\text{if} \quad y_i \left(\sum_{j=1}^{n} \alpha_j y_j K(x_i, x_j) + b^\lambda \right) > 1 \\ \alpha_i = C \quad &\text{if} \quad y_i \left(\sum_{j=1}^{n} \alpha_j y_j K(x_i, x_j) + b^\lambda \right) < 1 \\ \sum_i \alpha_i y_i \quad &= \quad 0 \end{aligned} \tag{5.32}$$

Interestingly, the above inequalities, which fully characterize the support vectors associated to the solution, are usually obtained as the Kuhn-Tucker conditions of the dual quadratic programming optimization problem [Vap98]. Looking at Eqs.(5.31-5.32), it is immediate to see that the box constraints ($0 \le \alpha_i \le C$) are due to the linearity of $\ell(yf(x))$ for $yf(x) < 1$, whereas sparsity ($\alpha_i = 0$) follows from the fact that $\ell(yf(x))$ is constant for $yf(x) > 1$.

5.6 Support Vector Algorithms as Regularization Networks

In this section we review various learning algorithms inspired by Support Vector Machines and show that each one defines a suitable regularization network. In particular the various

algorithms differ for the considered loss function and the way the offset term is treated. In the following we first give some comments on the role played by the loss functions and offset term, and then review various algorithms from a regularization network perspective.

We start observing that usually in the classification setting the loss function ℓ depends on its arguments through the product $yf(x)$, for this reason in the following we will use indifferently the expressions $\ell(y, f(x))$ and $\ell(yf(x))$. However we want to stress that this particular form of the loss function implicitly models situations in which false negative ($y = +1$ and $f(x) < 0$) and false positive ($y = -1$ and $f(x) > 0$) errors are equally penalized. More general situations have been considered in the literature (see for example [LLW02]), in the general case an extra factor depending on y has to be added to the loss function

$$\ell(y, f(x)) = L(y)\ell(yf(x)). \tag{5.33}$$

We will return on this last fact while considering one-class SVM. We can obtain different learning algorithms choosing different loss functions ℓ [EPP00]. Some choices we will consider in the following are

- the square loss $\ell(y, w) = (w - y)^2 = (1 - wy)^2$,
- the hinge loss $\ell(y, w) = \max\{1 - wy, 0\} =: |1 - wy|_+$,
- the truncated square loss $\ell(y, w) = \max\{1 - wy, 0\}^2 =: |1 - wy|_+^2$.

Moreover we note that we can modify (5.1) to include arbitrary penalization of the offset term, in particular let us consider the additive penalization b^c, with c a fixed parameter

$$\min_{f\in\mathcal{H}, b\in\mathbb{R}} \{\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i) + b) + \lambda(\|f\|_{\mathcal{H}}^2 \ + b^c)\}. \tag{5.34}$$

The case when the offset is not penalized is subsumed for $c = 0$. Moreover we will show in the following that both the cases $c = 1$ and $c = 2$ realize learning algorithms known in the literature. When considering RN algorithms one of the following cases usually occurs

1. no offset term is considered,
2. an unpenalized offset term is considered,
3. a quadratic penalization of the offset is added.

Interestingly case 3 is subsumed by case 1 as shown by the theorem below. We let $\mathcal{B} = \mathbb{R}$ be the space of constant functions which is a RKH space with kernel $K_{\mathcal{B}}(s, x) = 1$. Let

$s \in \mathcal{S}$, then it is well known (see [Aro50] or [Sch64]) that

$$\|s\|_{\mathcal{S}}^2 = \min_{\substack{(f,b)\in\mathcal{H}\times\mathbb{R} \\ \text{s.t. } f+b=s}} \{\|f\|_{\mathcal{H}}^2 + b^2\}. \tag{5.35}$$

Moreover the minimum is attained by the couple (f_s, b_s) given by $f_s = Qs$ and $b_s = s - Qs$. We are now ready to state the theorem

Theorem 18. *The couple $(\bar{f}, \bar{b})$ is a solution of the problem*

$$\min_{(f,b)\in\mathcal{H}\times\mathbb{R}} \{\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i) + b) + \lambda(\|f\|_{\mathcal{H}}^2 + b^2)\} \tag{5.36}$$

if and only if $\bar{f} = Q\bar{s}$ and $\bar{b} = \bar{s} - Q\bar{s}$, with $\bar{s}$ (that is $\bar{f} + \bar{b}$) solution of the problem

$$\min_{s\in\mathcal{S}} \{\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, s(x_i)) + \lambda \|s\|_{\mathcal{S}}^2\}. \tag{5.37}$$

Proof. Let us preliminarily notice that the minima achieved by the two functionals in the text are equal, in fact

$$\inf_{(f,b)\in\mathcal{H}\times\mathbb{R}} \{\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i) + b) + \lambda(\|f\|_{\mathcal{H}}^2 + b^2)\} \tag{5.38}$$

$$= \inf_{s\in\mathcal{S}} \inf_{\substack{(f,b)\in\mathcal{H}\times\mathbb{R} \\ \text{s.t. f+b=s}}} \{\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, s(x_i)) + \lambda(\|f\|_{\mathcal{H}}^2 + b^2)\} \tag{5.39}$$

$$= \inf_{s\in\mathcal{S}} \{\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, s(x_i)) + \lambda\|s\|_{\mathcal{S}}^2\}, \tag{5.40}$$

where the last equality descends from equation(5.35).

Now assume that $\bar{s}$ is a solution of problem(5.37), then

$$\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \bar{s}(x_i)) + \lambda(\|f_{\bar{s}}\|_{\mathcal{H}}^2 + (b_{\bar{s}})^2) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \bar{s}(x_i)) + \lambda\|\bar{s}\|_{\mathcal{S}}^2. \tag{5.41}$$

Recalling that we set $\bar{f} = Q\bar{s}$ and $\bar{b} = \bar{s} - Q\bar{s}$, it follows that the couple $(f_{\bar{s}}, b_{\bar{s}})$ is a solution of problem(5.36) since it attains the common minimum. This proves one half of the theorem.

On the other hand assume that $(\bar{f}, \bar{b})$ is a solution of problem(5.36), then setting $\bar{s} = \bar{f} + \bar{b}$ and recalling again equation(5.35), we get

$$\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \bar{s}(x_i)) + \lambda\|\bar{s}\|_{\mathcal{S}}^2 \leq \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \bar{f}(x_i) + \bar{b}) + \lambda(\|\bar{f}\|_{\mathcal{H}}^2 + \bar{b}^2). \tag{5.42}$$

But the r.h.s. of the inequality above is the common minimum of problems (5.37) and (5.36), then equality must hold. It follows both that $\bar{s}$ is a solution of problem (5.37) and that $(\bar{f}, \bar{b}) = (f_{\bar{s}}, b_{\bar{s}})$. This concludes the proof. □

The above results is a complement of theorem 17 corresponds simply to consider regularization with the kernel $K + 1$ and no-offset.

5.6.1 Non Standard SVM revisited

In this section we review a number of SVM-like algorithms and show that each one is indeed equivalent to a particular regularization network algorithm. that our analysis will not make use of the dual formulation of the algorithms. We add a few remarks before starting our discussion.

- the linear case $K(x, s) = x \cdot s$ has received much attention in machine learning literature (it realizes for example the original linear SVM). However we will not treat separately linear and non-linear cases, in fact we can always assume the existence of a mapping $\mathbf{\Phi}$ from X to a (possibly infinite dimensional) *feature space*, such that $K(x, s) = \mathbf{\Phi}(x) \cdot \mathbf{\Phi}(s)$ (see chapter 2),
- SVM algorithm in its primal formulation can be written as

$$\min_{f \in \mathcal{H}, b \in \mathbb{R}} \{\frac{1}{n} \sum_{i=1}^{n} |1 - y_i(f(x_i) + b)|_+ + \lambda \|f\|_{\mathcal{H}}^2\},$$

  but an alternative formulation equivalent to the previous one can be stated in terms of slack ξ variables and margin $\frac{1}{2}\mathbf{w} \cdot \mathbf{w}$ (in the feature space)

$$\min_{\mathbf{w}, b, \xi_i} \{C \sum_{i=1}^{n} \xi_i + \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle\}$$

  s.t. $\forall\ i$

$$y_i(\mathbf{w} \cdot \mathbf{\Phi}(x_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0,$$

- note that the relation between the parameter C in the standard slack variables formulation and the parameter λ in the regularization network formulation is simply $C = \frac{1}{2\lambda n}$,

5.6.1.1 L2-SVM

We consider (see [CST00] for reference) the problem

$$\min_{\mathbf{w},b,\xi_i}\{C\sum_{i=1}^{n}\xi_i^2+\frac{1}{2}\langle\mathbf{w},\mathbf{w}\rangle\}$$

s.t. $\forall\ i$

$$y_i(\mathbf{w}\cdot\mathbf{\Phi}(x_i)+b)\geq 1-\xi_i,\ \xi_i\geq 0,$$

that is we consider the squares of the slack variables and do not penalize the bias term b.

The above problem can be seen as the regularization network obtained considering the truncated square loss and penalizing the solutions by the seminorm $\|Q\cdot\|_{\mathcal{S}}$ in the sum space $\mathcal{S}$ defined in theorem 17. In fact we are simply considering the minimization problem

$$\min_{s\in\mathcal{H}}\{\frac{1}{n}\sum_{i=1}^{n}|1-y_is(x_i)|_+^2+\lambda\|Qs\|_{\mathcal{S}}^2\}.$$

We have seen that the above problem can be also written in the (more direct) form

$$\min_{f\in\mathcal{H},b\in\mathcal{B}}\{\frac{1}{n}\sum_{i=1}^{n}|1-y_i(f(x_i)+b)|_+^2+\lambda\|f\|_{\mathcal{H}}^2\}$$

5.6.1.2 Least Square SVM

We consider (see [CST00] and [SVGDB+02] for reference) the problem

$$\min_{\mathbf{w},b,\xi_i}\{C\sum_{i=1}^{n}\xi_i^2+\frac{1}{2}\langle\mathbf{w},\mathbf{w}\rangle\}$$

s.t. $\forall\ i$

$$y_i(\mathbf{w}\cdot\mathbf{\Phi}(x_i)+b)=1-\xi_i,\ \xi_i\geq 0,$$

the inequality constraints are now replaced by equality constraints and the square of the slack variables is considered, moreover an unpenalized offset term is added. It is easy to see that the above problem corresponds to a regularization network with square loss, the sum space $\mathcal{S}$ as hypothesis space and again penalty term $\lambda\|Qs\|_{\mathcal{S}}^2$. It follows that in this case the minimization problem can be written as

$$\min_{s\in\mathcal{S}}\{\frac{1}{n}\sum_{i=1}^{n}(1-y_is(x_i))^2+\lambda\|Qs\|_{\mathcal{S}}^2\}.$$

equivalently as

$$\min_{f\in\mathcal{H},b\in\mathcal{B}}\{\frac{1}{n}\sum_{i=1}^{n}(1-y_i(f(x_i)+b))^2+\lambda\left\|f\right\|_{\mathcal{H}}^2\}.$$

We now consider algorithms where the offset term is also penalized. For all the following algorithms the bias term appears explicitly in the form of the solution but it is now penalized by the complexity term.

5.6.1.3 Modified SVM

We consider (see [MM01] for reference) the problem

$$\min_{\mathbf{w},b,\xi_i}\{C\sum_{i=1}^{n}\xi_i+\frac{1}{2}(\langle\mathbf{w},\mathbf{w}\rangle+b^2)\}$$

s.t. $\forall\ i$

$$y_i(\mathbf{w}\cdot\mathbf{\Phi}(x_i)+b)\geq 1-\xi_i,\ \xi_i\geq 0,$$

that is, we now penalize the bias term b. The above problem can be easily interpreted in the sum space $\mathcal{S}$ with kernel $K+1$. In fact by theorem 18, it is equivalent to the problem

$$\min_{s\in\mathcal{S}}\{\frac{1}{n}\sum_{i=1}^{n}\left|1-y_i s(x_i)\right|_+ +\lambda\left\|s\right\|_{\mathcal{S}}^2\}.$$

As we mentioned before the above situation is formally equivalent to the case in which no offset term is considered.

5.6.1.4 Smooth SVM

We consider (see [LM01] for reference) the problem

$$\min_{\mathbf{w},b,\xi_i}\{C\sum_{i=1}^{n}\xi_i^2+\frac{1}{2}(\langle\mathbf{w},\mathbf{w}\rangle+b^2)\}$$

s.t. $\forall\ i$

$$y_i(\mathbf{w}\cdot\mathbf{\Phi}(x_i)+b)\geq 1-\xi_i,\ \xi_i\geq 0$$

that is we consider the square of the slack variables and do penalize the bias term b.

The above problem can be seen as the regularization network obtained considering the truncated square loss and searching for a solution in the sum space $\mathcal{S}$. In fact we are simply considering the minimization problem

$$\min_{s\in\mathcal{S}}\{\frac{1}{n}\sum_{i=1}^{n}|1-y_i s(x_i)|_+^2+\lambda\|s\|_{\mathcal{S}}^2\}.$$

5.6.1.5 Proximal SVM

We consider (see [FM01] for reference) the problem

$$\min_{\mathbf{w},b,\xi_i}\{C\sum_{i=1}^{n}\xi_i^2+\frac{1}{2}(\langle\mathbf{w},\mathbf{w}\rangle+b^2)\}$$

s.t. $\forall\ i$

$$y_i(\mathbf{w}\cdot\mathbf{\Phi}(x_i)+b)=1-\xi_i,\ \xi_i\geq 0$$

the inequality constraints are now replaced by equality constraints and the squares of the slack variables are considered, moreover the bias term b is penalized.

The above problem can be easily interpreted in the sum space $\mathcal{S}$ where the kernel is $K+1$. In fact we are simply considering the minimization problem

$$\min_{s\in\mathcal{S}}\{\frac{1}{n}\sum_{i=1}^{n}(1-y_i s(x_i))^2+\lambda\|s\|_{\mathcal{S}}^2\}.$$

It should be apparent that the above problem is completely equivalent to the regularized least squares algorithm in the RKH space $\mathcal{S}$.

5.6.1.6 One-class Support Vector machines as a regularization network

One-class classification techniques were originally developed to cope with binary classification problems in which statistics for one of the two classes was virtually absent ([TD99, SPST$^+$01]). In this setting the component of the training set $(x_i,y_i)_{i=1}^n$ labeled according to the minority class ($y=-1$ in the following) is intentionally removed, generating the reduced one-class training set $(x_i)_{i=1}^{n_+}$.

Intuitively the idea behind one-class SVM algorithm, in its simplest formulation, is looking for the smallest sphere enclosing the examples in the data space. Hence the training

procedure amounts to the solution of the following constrained minimization problem with respect to the balls of center $\mathbf{a}$ and radius R in the input space

$$\min_{R,\xi_i}\{C\sum_{i=1}^{n_+}\xi_i + R^2\}, \tag{5.43}$$

conditioned to the existence of a vector $\mathbf{a} \in \mathbb{R}^n$ s.t. $\forall\ i \leq n_+$

$$(x_i - \mathbf{a}) \cdot (x_i - \mathbf{a})^T \leq R^2 + \xi_i,\ \xi_i \geq 0.$$

A non-linear extension of the previous algorithm can be directly achieved by substituting scalar products with kernel functions. From a more geometrical point of view we consider balls in a suitable feature space, that is the squared distance appearing in (5.43) is now replaced by the expression $(\mathbf{\Phi}(x_i) - \mathbf{a}) \cdot (\mathbf{\Phi}(x_i) - \mathbf{a})^T$.

It can be shown (see for example [CS02b]) that the centers $\mathbf{a}$ can be mapped one to one with the functions f of the RKH space $\mathcal{H}$ of kernel $K(x, s) = \mathbf{\Phi}(x) \cdot \mathbf{\Phi}(s)$. This correspondence is such that $\mathbf{\Phi}(x) \cdot \mathbf{a} = f(x)$, so that the problem can be written as follows

$$\min_{R^2,\xi_i}\{C\sum_{i=1}^{n_+}\xi_i + R^2\}, \tag{5.44}$$

requiring that there exists a vector $f \in \mathcal{H}$ s.t. $\forall\ i \leq n_+$

$$K(x_i, x_i) - 2f(x_i) + \|f\|^2_{\mathcal{H}} \leq R^2 + \xi_i,\ \xi_i \geq 0.$$

In the minimization problems above the relevant complexity measure R^2 runs over non negative real values. In order to interpret the above problems from a regularization point of view it results convenient slightly modify the problem allowing the complexity measure run over unconstrained reals. This is obtained simply by introducing the real variable ρ and formulating the modified problem as follows

$$\min_{\rho,\xi_i}\{C\sum_{i=1}^{n_+}\xi_i + \rho\}, \tag{5.45}$$

conditioned to the existence of a function $f \in \mathcal{H}$ s.t. $\forall\ i \leq n_+$

$$K(x_i, x_i) - 2f(x_i) + \|f\|^2_{\mathcal{H}} \leq \rho + \xi_i,\ \xi_i \geq 0.$$

The two problems are indeed virtually equivalent, in fact it is straightforward to verify that whenever $C > \frac{1}{n_+}$ problems (5.44) and (5.46) have the same solutions. On the other hand if $C < \frac{1}{n_+}$ both problems are trivial: any solution of (5.44) attains $R = 0$ while no solution of (5.46) exists.

Introducing the offset b and the fixed bias function $\mathcal{D}(x)$, and suitably rescaling the parameter C and the kernel K, the previous problem becomes

$$\min_{f,b,\xi_i}\{\widetilde{C}\sum_{i=1}^{n_+}\xi_i + \frac{1}{2}(\|f\|^2_{\widetilde{\mathcal{H}}} + b)\}, \tag{5.46}$$

s.t. $\forall\ i \leq n_+$

$$\mathcal{D}(x_i) + f(x_i) + b \geq -\xi_i,\ \xi_i \geq 0.,$$

where we set

$$\begin{aligned} b &= \frac{1}{2}(\rho - \|f\|^2_{\mathcal{H}} - K(0,0)), \\ \mathcal{D}(x) &= \frac{1}{2}(K(x,x) - K(0,0)), \\ \widetilde{C} &= \frac{1}{4}C, \\ \widetilde{K} &= 2K. \end{aligned}$$

The couple (f, b) in (5.46) runs over $\widetilde{\mathcal{H}} \times \mathbb{R}$, with $\widetilde{\mathcal{H}}$ the RKH space of kernel $\widetilde{K}$.

Finally, by standard reasoning the problem can be rewritten in terms of loss function, penalty term and original two-class training set, in fact considering the loss function

$$\ell(y, w) = \theta(y) \cdot |-wy|_+,$$

which matches the general form in equation(5.33), we easily obtain

$$\min_{f\in\widetilde{\mathcal{H}},b\in\mathbb{R}}\{\frac{1}{n}\sum_{i=1}^{n}\ell(y_i, \mathcal{D}(x_i) + f(x_i) + b) + \lambda(\|f\|^2_{\widetilde{\mathcal{H}}} + b)\}. \tag{5.47}$$

By definition the fixed bias function $\mathcal{D}(x)$ is null for translation invariant kernel functions (e.g. the Gaussian kernel). In this case problem (5.47) fits the general regularization network form (5.34), with $c = 1$.

5.7 Consistency of Tikhonov Regularization with Convex Loss

In this section we discuss consistency of regularization networks defined by (5.1). The idea is to show a novel approach to the error analysis which is reminding of that we used for the square loss.

Throughout this section we make the following assumptions.

- **Loss functions**. We assume that the loss are admissible according to definition 1 and moreover we assume that there exists a constant C_0 such that, $\forall y \in Y$,

$$\ell(y, 0) \leq C_0. \tag{5.48}$$

  We recall that since the loss are admissible they are locally Lipschitz.

- **Noise assumption**. The output space is bounded, that is

$$Y = [-M, M], \qquad M > 0. \tag{5.49}$$

- **Kernel assumption**. The kernel satisfies assumption (1) in particular we recall that $\kappa = \sup_{x \in X} \sqrt{K(x, x)} \leq \infty$.

For sake of simplicity we do not take into account the bias term. We let $f_{\mathbf{z}}^{\lambda}$ be the solution of (5.1) and f^{λ} as the solution of problem (5.2) for $\mathcal{B} = \{\emptyset\}$.
Recalling that each admissible loss function induces a suitable target function $t_\rho \in \mathcal{F} = L^p(X, \rho_X)$, it is interesting to consider the following error decomposition for any $\lambda > 0$

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda}) - \mathcal{E}(t_\rho) \leq \underbrace{\mathcal{E}(f_{\mathbf{z}}^{\lambda}) - \mathcal{E}(f^{\lambda})}_{\text{sample error}} + \underbrace{\mathcal{E}(f^{\lambda}) + \lambda \left\| f^{\lambda} \right\|_{\mathcal{H}}^2 - \inf_{f \in \mathcal{H}} \mathcal{E}(f)}_{\text{approximation error}} + \underbrace{\inf_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}(t_\rho)}_{\text{irreducible error}}. \tag{5.50}$$

As usual the last term accounts for the fact that once we choose the hypotheses space $\mathcal{H}$ the best we can aim to is the minimum error in such a space. If $\mathcal{H}$ is dense in $L^2(X, \rho_X)$ such error is zero. Moreover the sample error is the component due to finite sampling and the approximation error is the error component due to the chosen regularization level, encoded by the value for the parameter λ.

It is useful to note some facts. No lower bounds are available for learning with convex loss. This largely depends on the fact that little is known about the behavior of the approximation error for functions other than the square loss. We recall that the approximation error goes to zero as λ goes to zero but it is not clear how to derive rates. In particular it is

often assumed that $\mathcal{E}(f^\lambda) + \lambda \left\| f^\lambda \right\|_{\mathcal{H}}^2 - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \leq C\lambda^s$ for some $C, s > 0$ but it is not known which class of target functions allows this kind of estimates. Some discussion on this issue can be found in [WYZ04]. As for the sample error various bounds are available and an overview of various approaches is given in [Ste04] for classification but the same techniques can be used for regression. Usually the following decomposition is considered

$$\mathcal{E}(f_{\mathbf{z}}^\lambda) - \mathcal{E}(f^\lambda) \leq \mathcal{E}(f_{\mathbf{z}}^\lambda) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^\lambda) + \mathcal{E}_{\mathbf{z}}(f^\lambda) - \mathcal{E}(f^\lambda) \tag{5.51}$$

and problem (5.1) and (5.2) are rewritten as constrained minimization problems so that results for ERM can be used. As we discussed in remark 3 when considering this approach one should be careful on the dependence to the regularization parameter λ. A more complete list of references can be found [WYZ04, Ste04], here we just recall the various term in (5.51) can be controlled via the stability approach [BE02] or the complexity based approaches - for example those using Rademacher complexities [BM02] or covering numbers [Ste04].

Given the above premises we present our results about consistency of regularization networks. The main goal is to present an approach to the analysis of the sample error different to those we just mentioned which does not make use of any complexity measure. Given such a bound we can derive a sufficient condition for choosing λ in such a way that consistency is ensured. The following theorem summarizes such a result.

Theorem 19 (Consistency). *Let $f_{\mathbf{z}}^\lambda$, be the solution of problems (5.1). We assume the loss ℓ to be admissible and denote with L_λ the Lipschitz constant of $\ell(y, w)$ for $w \in [-\frac{C_0}{\sqrt{\lambda}}, \frac{C_0}{\sqrt{\lambda}}]$. Moreover we assume the kernel to satisfy assumption 1 and $Y = [-M, M]$. Then if we choose $\lambda_n = \lambda(n)$ such that*

$$\lim_{n\to 0} \frac{L_{\lambda_n}}{\lambda_n \sqrt{n}}$$

and let $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$ then

$$\lim_{n\to\infty} \Pr\left(\mathcal{E}(f_{\mathbf{z}}) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \right) = 0.$$

The proof of the above result relies on estimates for sample and approximation error. In particular for the sample error we consider the following line of reasoning. For admissible loss functions we can relate the expected error to the deviation of $f_{\mathbf{z}}^\lambda$ and f^λ in the $\mathcal{H}$-norm, in fact it can be shown that

$$\begin{aligned} |\mathcal{E}(f_{\mathbf{z}}^\lambda) - \mathcal{E}(f^\lambda)| &\leq \int_{X\times Y} |\ell(y, f_{\mathbf{z}}^\lambda(x)) - \ell(y, f^\lambda(x))| d\rho(x, y) \\ &\leq L_\lambda \int_{X\times Y} |f_{\mathbf{z}}^\lambda(x) - f^\lambda(x)| d\rho(x, y) \\ &\leq L_\lambda \left\| f_{\mathbf{z}}^\lambda - f^\lambda \right\|_\infty \leq \kappa L_\lambda \left\| f_{\mathbf{z}}^\lambda - f^\lambda \right\|_{\mathcal{H}} \end{aligned} \tag{5.52}$$

for a suitable constant L_λ depending on the regularization parameter λ. Now recalling that from theorem 16

$$\begin{aligned} f^\lambda_{\mathbf{z}}(x) &= -\frac{1}{2\lambda}\sum_{i=1}^n \alpha_i K(x,x_i), \qquad \alpha_i \in \partial\ell(y_i, f^\lambda_{\mathbf{z}}(x_i)) \\ f^\lambda(x) &= -\frac{1}{2\lambda}\int_{X\times Y} \alpha(s,y)K(x,s)d\rho(s,y), \qquad \alpha(x,y) \in \partial\ell(y, f^\lambda(x)) \end{aligned}$$

the idea is to look at $f^\lambda_{\mathbf{z}}$ and f^λ as the empirical and expected average of a $\mathcal{H}$-valued random variable and use concentration inequality for Hilbert space valued random variables. The main difficulty is that the α_i are different to $\alpha(x_i, y_i)$ and we have to relate them in some way. Indeed this can be achieved via the following lemma (see [Ste03, CS05]).

Lemma 8. *Let $f^\lambda_{\mathbf{z}}$, f^λ be the solution of problems (5.1) and problem (5.2) (for $\mathcal{B} = \{\emptyset\}$). Under the same assumption of theorem 16 we can prove that*

$$\left\|f^\lambda_{\mathbf{z}} - f^\lambda\right\|_{\mathcal{H}} \leq \frac{1}{\lambda}\left\|\frac{1}{n}\sum_{i=1}^n \alpha(x_i,y_i)K_{x_i} - \int_{X\times Y}\alpha(x,y)K_x d\rho(x,y)\right\|_{\mathcal{H}} \tag{5.53}$$

with $\alpha(x,y) \in \partial\ell(y, f^\lambda(x))$, $(x,y) \in X \times Y$ ρ-a.s.

Using (5.52) and (5.53) we get the following bound for the sample error (which is a simplified version of a theorem in [CS05]).

Theorem 20 (Sample Error). *Let $f^\lambda_{\mathbf{z}}$, f^λ be the solution of problems (5.1) and problem (5.2) (for $\mathcal{B} = \{\emptyset\}$). We assume the loss ℓ to be admissible and denote with L_λ the Lipschitz constant of $\ell(y,w)$ for $w \in [-\frac{C_0}{\sqrt{\lambda}}, \frac{C_0}{\sqrt{\lambda}}]$. Moreover we assume the kernel to satisfy assumption 1 and $Y = [-M, M]$. Then for $0 < \eta \leq 1$, $n \in \mathbb{N}$ and $\lambda > 0$ the following inequality holds with probability at least $1 - \eta$*

$$\mathcal{E}(f^\lambda_{\mathbf{z}}) - \mathcal{E}(f^\lambda) \leq \frac{\sqrt{2}L_\lambda \kappa}{\lambda\sqrt{n}}\sqrt{\log\frac{2}{\eta}} \tag{5.54}$$

We add the following remark.

Remark 25. *As we observed for the square loss, bounding the expected error via the bound for the $\mathcal{H}$-norm worsen the results since such norm is very strong. Moreover comparing with similar results for the squared loss here we loose a square exponent. This is because there is no metric naturally associated to expected error and we have to use the Lipschitz property to define one.*

For the approximation error we recall the following trivial result

Theorem 21 (Approximation Error). *If f^λ solves problem (5.2) for $\mathcal{B} = \{\emptyset\}$ then*

$$\lim_{\lambda \to 0} \mathcal{E}(f^\lambda) + \lambda \|f\lambda\|_{\mathcal{H}}^2 = \inf_{f \in \mathcal{H}} \mathcal{E}(f).$$

If we look at the inequality (5.54) we can see that choosing $\lambda_n = \lambda(n)$ such that

$$\lim_{n \to \infty} \frac{L_{\lambda_n}}{\lambda_n \sqrt{n}} = 0$$

we are sure that the sample error goes to zero as $n \to \infty$. Then the proof of theorem 19 follows noting that for the above choice, λ_n goes to zero as $n \to \infty$ so that the approximation error converge to zero.

5.7.1 Proofs

We first give the proof of lemma 8.

Proof of lemma 8. Recalling the definition of $\alpha(x, y)$ by definition of subgradient we have

$$\ell(y, f_{\mathbf{z}}^\lambda(x)) - \ell(y, f^\lambda(x)) \geq \alpha(x, y) \left\langle f_{\mathbf{z}}^\lambda - f^\lambda, K_x \right\rangle_{\mathcal{H}}.$$

If we now take the empirical average on the sample we get

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^\lambda) - \mathcal{E}_{\mathbf{z}}(f^\lambda) \geq \left\langle f_{\mathbf{z}}^\lambda - f^\lambda, \frac{1}{n} \sum_{i=1}^n \alpha(x_i, y_i) K x_i \right\rangle_{\mathcal{H}}. \tag{5.55}$$

Moreover we note that the following equality holds true

$$\lambda \left\| f_{\mathbf{z}}^\lambda \right\|_{\mathcal{H}}^2 - \lambda \left\| f^\lambda \right\|_{\mathcal{H}}^2 - \lambda \left\| f_{\mathbf{z}}^\lambda - f^\lambda \right\|_{\mathcal{H}}^2 = 2\lambda \left\langle f_{\mathbf{z}}^\lambda - f^\lambda, f^\lambda \right\rangle_{\mathcal{H}} \tag{5.56}$$

By definition of $f_{\mathbf{z}}^\lambda$ we have $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^\lambda) + \lambda \left\| f_{\mathbf{z}}^\lambda \right\|_{\mathcal{H}}^2 - \mathcal{E}_{\mathbf{z}}(f^\lambda(x)) - \lambda \left\| f^\lambda \right\|_{\mathcal{H}}^2 \leq 0$ so that if we now put (5.55) and (5.56) together we get the following inequality

$$\begin{aligned} \lambda \left\| f_{\mathbf{z}}^\lambda - f^\lambda \right\|_{\mathcal{H}}^2 &\leq \left\langle f^\lambda - f_{\mathbf{z}}^\lambda, \frac{1}{n} \sum_{i=1}^n \alpha(x_i, y_i) K(x, x_i) + 2\lambda f^\lambda \right\rangle_{\mathcal{H}} \\ &\leq \left\| f_{\mathbf{z}}^\lambda - f^\lambda \right\|_{\mathcal{H}} \left\| \frac{1}{n} \sum_{i=1}^n \alpha(x_i, y_i) K(x, x_i) - 2\lambda f^\lambda \right\|_{\mathcal{H}} \end{aligned}$$

where we used Schwartz inequality. The proof follows noting that

$$2\lambda f^\lambda = -\int_{X \times Y} \alpha(x, y) K_x d\rho(x, y)$$

by definition of f^λ. □

We now can prove the bound for the sample error.

Proof of theorem 20. To prove the theorem we note that $\left\|f_{\mathbf{z}}^{\lambda}\right\|_{\mathcal{H}}, \left\|f^{\lambda}\right\|_{\mathcal{H}} \leq \frac{C_0}{\sqrt{\lambda}}$, in fact

$$\lambda \left\|f^{\lambda}\right\|_{\mathcal{H}}^2 \leq \mathcal{E}(f^{\lambda}) + \lambda \left\|f^{\lambda}\right\|_{\mathcal{H}}^2 \leq \mathcal{E}(0) \leq C_0$$

and similarly one can prove the bound for $\left\|f_{\mathbf{z}}^{\lambda}\right\|_{\mathcal{H}}$. Since the loss $\ell(y, w)$ is admissible (and hence convex) it is Lipschitz in the interval $[-\frac{C_0}{\sqrt{\lambda}}, \frac{C_0}{\sqrt{\lambda}}]$ so that inequalities (5.52) hold true. Putting together (5.52) and (5.53) we get

$$|\mathcal{E}(f_{\mathbf{z}}^{\lambda}) - \mathcal{E}(f^{\lambda})| \leq \frac{1}{\lambda} \left\| \frac{1}{n} \sum_{i=1}^{n} \alpha(x_i, y_i) K_{x_i} - \int_{X \times Y} \alpha(x, y) K_x d\rho(x, y) \right\|_{\mathcal{H}}$$

If we define the random variable $\xi : Z \to \mathcal{H}$ defined by $\xi = \alpha(x, y) K_x$ then

$$\|\alpha(x, y) K_x\|_{\infty} \leq L_{\lambda} \kappa$$

since $\|\alpha(x, y)\|_{\infty} \leq L_{\lambda}$ by definition of α and the properties of the subgradient. Finally we can apply inequality (2) to get the result. □

Proof of theorem 21. The proof follows noting that if we let $f_{\varepsilon} \in \mathcal{H}$ such that $\mathcal{E}(f_{\varepsilon}) \leq \inf_{\mathcal{H}} \mathcal{E}(f) + \varepsilon$ then it exists λ_0 such that for $\lambda < \lambda_0$

$$\lambda \|f_{\varepsilon}\|_{\mathcal{H}}^2 \leq \varepsilon.$$

□

5.7.2 Bayes Consistency

We now consider the problem of Bayes consistency of regularization networks. Again the notion of convexity turns out to be crucial. We recall that in the context of classification the misclassification risk (4.55) is often considered as the natural error measure. It is useful to note that

$$R(f) = \Pr\left(\operatorname{sign} f(x) \neq y\right) = \int_{X \times Y} \theta(-y f(x)) d\rho(x, y),$$

that is the misclassification risk is the expected error induced by the loss $\theta(t) = \chi_{\{t>0\}}$ which is the Heavyide step-function. Clearly we would be tented to consider algorithms minimizing the empirical misclassification risk

$$\frac{1}{n} \sum_{i=1}^{n} \theta(-y_i f(x_i),$$

but it is known that this leads to computationally intractable problems. For this reason the misclassification loss $\theta(\cdot)$ is usually replaced by a convex loss function (sometimes called surrogate loss) which should be close to it.

As we consider algorithms based on convex loss function it is natural to wonder weather they allow us to achieve Bayes consistency, that is to recover the best classification error if enough examples are available. An exhaustive study on the subject can be found in [BJM05] where necessary and sufficient conditions are given on the loss functions. In the following we first discuss some sufficient conditions (see [RDVC$^+$04] and references in [WYZ04]) and then recall some of the results in [BJM05]. Recall that for a given estimator $f_{\mathbf{z}}$ the corresponding classification rule is obtained via thresholding, that is considering $\operatorname{sign} f_{\mathbf{z}}$. The idea is then to bound the excess risk

$$R(f_{\mathbf{z}}) - R(f_B), \tag{5.57}$$

using the bounds on the excess expected error. This requires comparison results (like those we discussed for the square loss) for general convex loss. Surprisingly this is particular easy for the hinge loss function since a direct calculation [RDVC$^+$04] shows that

$$t_\rho = f_B.$$

Given this premise we can simply check that for the hinge loss

$$R(f_{\mathbf{z}}) - R(f_B) \leq \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(t_\rho)$$

in fact $R(f_B) = \mathcal{E}(t_\rho)$ and moreover $R(f_{\mathbf{z}}) \leq \mathcal{E}(f_{\mathbf{z}})$ since the hinge loss upper bounds the misclassification loss. The following result [RDVC$^+$04] indicates the importance of the convexity assumption.

Proposition 10. *Assume that the loss function $\ell(y, w) = \ell(wy)$ is convex and that it is decreasing in a neighborhood of 0. Let t_ρ be the corresponding target function, if $t_\rho(x) \neq 0$, then*

$$f_B(x) = \mathit{sign} t_\rho(x).$$

We add few remarks before giving the proof.

Remark 26. *In the above proposition we used the fact that in classification the loss function $\ell(y, f(x))$ is in fact a function of the product $yf(x)$.*

Proof of proposition 10. We recall that, since ℓ is convex, ℓ admits left and right derivative in 0 and, since it is decreasing, $\ell'_-(0) \leq \ell'_+(0) < 0$. Observe that

$$\mathcal{E}(f) = \int_X \left(\rho(1|x)\ell(f(x) + (1 - \rho(1|x))\ell(-f(x))\right) d\rho_X(x),$$

and we have $t_\rho(x) = \operatorname{argmin}_{w\in\mathbb{R}} \psi(w)$, where

$$\psi(w) = \rho(1|x)\ell(w) + (1-\rho(1|x))\ell(-w)$$

(we assume existence and uniqueness to avoid pathological cases). Assume, for example, that $\rho(1|x) > \frac{1}{2}$. Then,

$$\begin{aligned}\psi'_-(0) &= \rho(1|x)\ell'_-(0) - (1-\rho(1|x))\ell'_+(0)\\ &\leq \rho(1|x)\ell'_+(0) - (1-\rho(1|x))\ell'_+(0)\\ &= (2\rho(1|x)-1)\ell'_+(0) \leq 0,\end{aligned}$$

Since ψ is also a convex function in w, this implies that for all $w \leq 0$

$$\psi(w) \geq \psi(0) + \psi'_-(0)w \geq \psi(0),$$

so that the minimum point w^* of $\psi(w)$ is such that $w^* \geq 0$. Since $t_\rho(x) = w^*$, it follows that if $t_\rho(x) \neq 0$

$$\operatorname{sign} t_\rho(x) = \operatorname{sign}(2\rho(1|x)-1) = f_B(x).$$

This ends the proof. □

Remark 27. *The technical condition $t_\rho(x) \neq 0$ is satisfied by all the loss functions we discussed and is equivalent to require the differentiability of ℓ in the origin. In fact consider the case $\rho(1|x) > \frac{1}{2}$. Computing the right derivative of ψ in 0, $\psi'_+(0)$, and observing that $\psi'_+(0) \geq 0$ for $\rho(1|x) \in (\frac{1}{2}, \frac{\ell'_-(0)}{\ell'_-(0)+\ell'_+(0)})$, it follows that this interval is empty if and only if $\ell'_-(0) = \ell'_+(0)$.*

In [WYZ04] the following comparison result for general convex loss function is given.

Theorem 22. *Assume that the loss function $\ell(t) = \ell(yw)$ is convex, differentiable at zero with $\ell'(0) < 0$ and $\inf_{t\in\mathbb{R}} \ell(t) = 0$. Then for any measurable function $f \in \mathbb{R}^X$ and probability measure ρ on $X \times Y$*

$$R(f_{\mathbf{z}}) - R(f_B) \leq \sqrt{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(t_\rho)}.$$

Finally [BJM05] a careful analysis is provided which highlights the different behavior of the various loss. Its proof relies on a notion of convex transform, that we denote with τ, whose precise definition is out of our scopes. Here we note that from the properties of τ, it follows that for any sequence $\theta_i \in [0,1]$

$$\tau(\theta_i) \to 0 \text{ if and only if } \theta_i \to 0. \tag{5.58}$$

In particular τ can be explicitly computed for various loss.

- hinge: $\tau(\theta) = \frac{\theta}{2}$,
- square: $\tau(\theta) = \theta^2$,
- truncated square: $\tau(\theta) = \theta^2$.

The following theorem gives the general comparison result for convex loss.

Theorem 23. *For any nonnegative loss function $\ell(yf(x))$, measurable function $f \in \mathbb{R}^X$ and probability measure ρ on $X \times Y$*

$$\tau(R(f_{\mathbf{z}}) - R(f_B)) \leq \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(t_\rho).$$

The above results can be improved if Tsybakov's noise condition (see section 4.5) is assumed and we refer to [BJM05] for further details. We end this section with a few comments. First, indeed in classification we always work with real valued functions that we have to threshold to obtain classification rule. Second, to obtain risks bounds we need to derive suitable comparison results and then study the error decomposition (**??**). This shows in particular that we cannot hope to have Bayes consistency unless the hypotheses space is dense in the target space $\mathcal{F}$.

Chapter 6

Conclusion

In this thesis we proposed an approach to learning whose goal is to emphasize and exploit the connection with ill-posed inverse problems. In particular we considered the question of whether the learning problem can be written as an inverse problem and give a positive answer for quadratic loss and hypotheses spaces that are reproducing kernel Hilbert spaces. Though there are differences with standard inverse problems, for examples new perturbation measures have to be defined, we can import many results and concepts. In particular we can define new kernel methods whose theoretical properties can studied in an elegant, unified frameworks. For such algorithms we can provide fast convergence rate both for regression and classification. When considering more general loss functions there is no natural way to define learning as an inverse problem. Nonetheless if we assume the loss to be convex we can use again a functional analytical approach using tools from convex analysis. The latter allows to study in a unified way the properties of algorithms inspired by Tikhonov regularization, i.e. regularization networks. Indeed we can completely characterize existence, uniqueness and explicit form of the regularized solutions giving new quantitative version of the representer theorem. In particular we take explicit care of the presence of an unpenalized off-set term investigating the relation with regularization with a penalty which is a semi-norm. Using the fact that the representer result holds both for the sample and the population case we can give an easy prove of consistency which is a simplified version of a results in [CS05].

We conclude discussing some open problems and directions for future developments. A main aspect of our analysis is that it provides a abstract definition of regularization with the square loss and gives a general framework to describe a large class of algorithms. Nonetheless we have a separate analysis for the case when the model exists and when the model does not exists. In the latter case we have to impose extra condition on the regularization and it would be interesting to investigate further if they can be dropped. In particular it is not clear if the results now available for Tikhonov regularization can be

recovered and extended to other algorithms in a unified framework. Moreover in this case we considered stronger noise assumption which should be possible to relax. In any case our study did not investigate the impact of the choice of the kernel. Indeed a more substantial study in this direction is given in [CDV05b] for Tikhonov regularization and it is not clear if such results can be extended to other algorithms.
More generally it would be interesting to consider other kind of regularization such as sparsity enhancing regularization and regularization with differential operators.

Many open problems are still to be studied when considering more general loss functions. Though many algorithms exist using various kind of loss functions their properties are not always clear. For example few results exist considering their approximation properties. The non linear structure of the problems for general loss makes it cumbersome to extend our connection with inverse problems (see discussion in [DVRCG04]). Nonetheless it would be interesting to see if regularization for some large class of algorithms can be described in abstract way as we did for the square loss.

Concluding, we note that one of the problem that motivated our study is far from being solved: the choice of the regularization paramter. The theory often proposes methods to choose the parameters which are asymptotically correct but not usable in practice. Moreover they often requires to know in advance the smoothness of the target function. In practice heuristics based on hold-out or cross validation are often used whose theoretical properties are not clear. Both from the theoretical and the practical point of view it would be desirable to define data driven, adaptive parameter choice with theoretical guarantees.

Appendix A

Some Mathematical tools

A.1 Convergence of Random Variables and Concentration Inequalities

We collect a few useful definitions and results. We refer to [Dud02] for details.

A **Polish space** is a separable completely metrisable topological space; that is, a space homeomorphic to a complete metric space that has a countable dense subset.

Recall that a **probability space** is a triple $(\Omega, \mathcal{B}, \mu)$ such that:
• Ω is a non empty set often called sample space,
• $\mathcal{B}$ is a σ-algebra of subsets of Ω whose elements are called "events". To say that $\mathcal{B}$ is a σ-algebra implies per definition that it contains Ω, that the complement of any event is an event, and that the union of any (finite or countably infinite) sequence of events is an event; • μ is a probability measure on $\mathcal{B}$, i.e. a measure such that $\mu(\Omega) = 1$.
A **measurable space** is simply a set Ω endowed with a σ-algebra $\mathcal{B}$.

A **random variable** is defined as a measurable function from a probability space to some measurable space. This measurable space is the space of possible values of the variable, for example we can think of it as the space of real numbers with the Borel σ-algebra. We always write

$$\mu(X \geq 0)$$

meaning

$$\mu(\{\omega \in \Omega | X(\omega) \geq 0\}).$$

We now list several notion of convergences for random variables. In the following, for $n \in \mathbb{N}$, we let X_n, X be real-valued random variables on a probability space $(\Omega, \mathcal{B}, \mu)$.

Convergence in probability, we say that the sequence X_n converges toward X in probability if

$$\lim_{n\to\infty} \mu\left(|X_n - X| \geq \varepsilon\right) = 0$$

for every $\varepsilon > 0$.
Convergence almost surely, we say that the sequence X_n converges almost surely or almost everywhere or with probability 1 or strongly to-wards X if

$$\mu\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Convergence in expectation, we say that the sequence X_n converges in expectation to-wards X if

$$\lim_{n\to\infty} \mathrm{E}\left(|X_n - X|\right) = 0$$

where E denotes the expectation and $\mathrm{E}\left(X_n\right) \leq \infty$. Here we recall informally some relations between the above notions of convergence. The following implications hold true.

- Convergence in expectation implies convergence in probability. This can be easily seen from Markov's inequality (see below). If the sequence of random variable is bounded almost surely, then convergence in probability implies convergence in expectation.
- Almost sure convergence implies convergence in probability. To invert the last result we need some extra condition as it is described by Borel-Cantelli lemma. The latter says that convergence in probability implies almost sure convergence if for all $\varepsilon > 0$,

$$\sum_n \mu\left(|X_n - X| > \epsilon\right) < \infty,$$

Finally we recall some useful probabilistic inequalities giving concentration results for random variables,

- **Markov's inequality**. If X is a real-valued random variable and $\varepsilon > 0$, then

$$\mu\left(|X| \geq \varepsilon\right) \leq \frac{\mathrm{E}(|X|)}{\varepsilon}.$$

- **Hoeffding's inequality**. Let $X_1, \ldots, X_n$ be a sequence of independent random variables such that X_i falls in the interval $[a_i, b_i]$ with probability one. Then if we let

$$Z = \frac{1}{n}\sum_{i=1}^{n} X_i$$

we have for all $\varepsilon > 0$

$$\mu\left(Z - \mathrm{E}[Z] \geq \varepsilon\right) \leq e^{-\frac{2n\varepsilon^2}{\Sigma(b_i - a_i)^2}}$$

Finally we discuss some relations between the different notions of convergence for real valued random variables.

A.2 Linear Operators and Spectral Theory

We refer to [Lan93] or [Rud91] for details.

An **Hilbert space** is linear space endowed with a scalar product and which is complete with respect to the induced norm. Completeness means that every Cauchy sequence of elements of the space converges in norm to an element in the space. For a Hilbert space $\mathcal{H}$ we denote with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, $\|\cdot\|_{\mathcal{H}}$ the scalar product and the norm respectively. A **Banach space** is a complete normed linear space.

We recall that a **bounded linear operator** A is a linear transformation between normed linear spaces $\mathcal{H}$ and $\mathcal{G}$ for which the ratio of the norm of Af to that of f, is bounded by the same number, $\forall f \in \mathcal{G}$. In other words, there exists some $C > 0$ such that for all $f \in \mathcal{G}$,

$$\|Af\| \leq C \|f\|_{\mathcal{H}}.$$

The smallest of such C is called the operator norm $\|A\|$ of A. We denote with $Im(A)$ the range of A. We denote with $\mathcal{B}(\mathcal{H}, \mathcal{G})$ the Banach space of bounded linear operators from $\mathcal{H}$ to $\mathcal{G}$ endowed with the operator norm $\|\cdot\|$. We write $\mathcal{B}(\mathcal{H})$ for the operators mapping $\mathcal{H}$ into itself. If $\mathcal{H}, \mathcal{G}$ are Hilbert spaces we denote with A^* the unique operator such that $\langle Af, g \rangle_{\mathcal{G}} = \langle f, A^* g \rangle_{\mathcal{H}}$, $f \in \mathcal{H}$ $g \in \mathcal{G}$; that is the adjoint operator. An operator is self-adjoint if $A = A^*$.

We often use the **polar decomposition** for bounded linear operator between Hilbert spaces. The latter is a canonical factorization of a bounded linear operator as the product of a partial isometry and a non-negative self-adjoint operator. More precisely each linear bounded operator (between Hilbert spaces) admits a polar decomposition $A = U|A|$ where $U : \mathcal{H} \to \mathcal{G}$ is a partial isometry and $|A| = \sqrt{A^* A}$ the unique operator such that $|A|^2 = A^* A$.

A **compact operator** is a linear operator A from a Banach space $\mathcal{H}$ to another Banach space $\mathcal{G}$, such that the image under A of any bounded subset of $\mathcal{H}$ is a relatively compact subset of $\mathcal{G}$. Recall that a relatively compact subspace of a space is a subset whose closure is compact.

A **Hilbert-Schmidt operator** is a bounded operator A on a Hilbert space $\mathcal{H}$ such that there exists an orthonormal basis $\{e_i : i \in I\}$ of $\mathcal{H}$ with the property

$$\sum_{i \in I} \|Ae_i\|^2 < \infty.$$

If this is true for one orthonormal basis, it is true for any other orthonormal basis. The space of Hilbert-Schmidt operators is the an Hilbert space endowed the Hilbert-Schmidt inner product defined as

$$\langle A, B\rangle_2 = \sum_{i\in I}\langle Ae_i, Be_i\rangle_{\mathcal{G}}.$$

This definition is independent of the choice of orthonormal basis. We denote with $\mathcal{B}_2(\mathcal{H},\mathcal{G})$ (or $\mathcal{B}_2(\mathcal{H})$) the Hilbert space of Hilbert-Schimdt operators.

A **trace class operator** is a bounded linear operator A over a Hilbert space $\mathcal{H}$ such that for some (and hence all) orthonormal bases $\{e_i : i \in I\}$ of $\mathcal{H}$ the sum of positive terms

$$\sum_{i\in I}\langle\sqrt{A^*A}\, e_i, e_i\rangle$$

is finite. In this case, the sum

$$\sum_{i\in I}\langle Ae_i, e_i\rangle$$

is absolutely convergent and is independent of the choice of the orthonormal basis. This value is called the trace of A, denoted by $\mathrm{Tr}(A)$. We denote with $\mathcal{B}_1(\mathcal{H},\mathcal{G})$ (or $\mathcal{B}_1(\mathcal{H})$) the space of Hilbert-Schimdt operators.
We note that $\mathcal{B}_2(\mathcal{H},\mathcal{G}) \subset \mathcal{B}_1(\mathcal{H},\mathcal{G})$ and they are both contained in the space of compact operators.

We end this section recalling some notion of spectral calculus. This is useful since we will need the notion of operator function. For sake of simplicity we just consider compact operators.

For an operator A the **spectrum** is the set of all the scalar values λ, the eigenvalues, such that $A + \lambda$ is not invertible. For compact self-adjoint operators this set is discrete. An **eigensystem** (σ_i, v_i) consists of all non-zero eigenvalues σ_i and the corresponding orthonormal basis of eigenvectors v_i such that $Av_i = \sigma_i v_i$. For such a basis we can write

$$Af = \sum_{i=1}^{\infty}\sigma_i\,\langle f, v_i\rangle_{\mathcal{H}}\, v_i.$$

If A is not self-adjoint we might have that the spectrum is the empty set. Nonetheless we can define the **singular system** $(\sigma_i; v_i, u_i)$ where σ_i are the eigenvalues of A^*A (and also AA^*) and the eigenvectors v_i and u_i satisfy the following equations

$$\begin{aligned} Av_i &= \sigma_i u_i, \quad Af = \sum_{i=1}^{\infty}\sigma_i\,\langle f, v_i\rangle_{\mathcal{H}}\, u_i,\ f\in\mathcal{H} \\ A^*u_i &= \sigma_i v_i \quad A^*g = \sum_{i=1}^{\infty}\sigma_i\,\langle g, u_i\rangle_{\mathcal{H}}\, v_i,\ g\in\mathcal{G} \end{aligned}$$

We make frequent use of the following facts:

- For a compact self-adjoint operator we can evaluate the operator norm observing that
$$\|A\| = \sup_{i \in I} \sigma_i;$$
- We can define an operator function $h(A)$ meaning
$$h(A) = \sum_{i=1}^{\infty} h(\sigma_i) \langle f, v_i \rangle_{\mathcal{H}} v,$$
clearly
$$\|h(A)\| = \sup_{\sigma_i \in I} h(\sigma_i).$$

A.3 Convex Functions in Infinite Dimensional Spaces

The proof of Theorem 16 is based on the properties of convex functions defined on infinite dimensional spaces. In particular, we use the notion of subgradient that extends the notion of derivative to convex non-differentiable functions. In this appendix we collect the results we need. For details see the book [ET83b] and also [ET74].

Let $\mathcal{H}$ be a Banach space and $\mathcal{H}^*$ its dual. A function $F : \mathcal{H} \to \mathbb{R}$ is *convex* if

$$F(tv + (1-t)w) \leq tF(v) + (1-t)F(w),$$

for all $v, w \in \mathcal{H}$ and $t \in [0,1]$ (if the strict inequality holds for $t \in (0,1)$, F is called *strictly convex*).

Let $v_0 \in \mathcal{H}$ such that $F(v_0) < +\infty$. The *subgradient* of F at point $v_0 \in \mathcal{H}$ is the subset of $\mathcal{H}^*$ given by

$$\partial F(v_0) = \{w \in \mathcal{H}^* \,|\, F(v) \geq F(v_0) + \langle w, v - v_0 \rangle \,,\ \forall v \in \mathcal{H}\}. \tag{A.1}$$

where $\langle \cdot, \cdot \rangle$ is the pairing between $\mathcal{H}^*$ and $\mathcal{H}$. If $F(v) = +\infty$, we let $\partial F(v_0) = \emptyset$.

In the following proposition we summarize the main properties of the subgradient we need.

Proposition 11. *The following facts hold:*

1. *If F is differentiable at v_0, the subgradient reduces to the usual gradient $F'(v_0)$.*

2. *If F is defined on $\mathbb{R}$ and $F(v_0) < +\infty$, then F admits left and right derivative and*

$$\partial F(v_0) = [F'_-(v_0), F'_+(v_0)].$$

3. *Assume that $F \neq +\infty$. A point v_0 is a minimizer of F if and only if $0 \in \partial F(v_0)$.*

4. *If F is continuous and*

$$\lim_{\|v\|_{\mathcal{H}} \to +\infty} F(v) = +\infty.$$

then F has a minimizer. If F is strictly convex, the minimizer is unique.

5. *Let G be another convex function on $\mathcal{H}$. Assume that there is $v_0 \in \mathcal{H}$ such that F and G are continuous and finite at v_0. Let $a, b \geq 0$, then $aF + bG$ is convex and, for all $v \in \mathcal{H}$,*

$$\partial(aF + bG)(v) = a(\partial F)(v) + b(\partial G)(v).$$

6. *Let $\mathcal{H}'$ be another Banach space and J be a continuous linear operator from $\mathcal{H}'$ into $\mathcal{H}$. Assume that there is $v'_0 \in \mathcal{H}'$ such that F is continuous and finite at Jv'_0. For all $v' \in \mathcal{H}'$*

$$(\partial F \circ J)(v') = J^*(\partial F)(Jv'),$$

where $J^* : \mathcal{H}^* \to \mathcal{H}'^*$ is the adjoint of J defined by

$$\langle v', J^* v \rangle_{\mathcal{H}'} = \langle Jv', v \rangle_{\mathcal{H}}.$$

for all $v \in \mathcal{H}$ and $v' \in \mathcal{H}'$.

Proof. We simply give the references to the book of [ET83b].

1. Prop. III.2.8
2. Prop. III.2.7
3. It is a simple consequence of Prop. III.3.1
4. It is a simple consequence of Prop. II.4.6.
5. Prop. III.2.13
6. Prop. III.2.12

□

We now recall the definition of *Nemitski* functional, adapted to our framework [**?**,]p.143]ekeland. Let Z be a locally compact second countable space, ρ be a finite measure on Z, and $W : Z \times \mathbb{R} \to [0, +\infty[$ be a measurable function on $Z \times \mathbb{R}$ such that $W(z, \cdot)$ is convex for all $z \in Z$ (since $W(z, \cdot)$ is convex on $\mathbb{R}$, it is continuous).

Let $p \in [1, +\infty[$ and $L^p(Z, \rho)$ be the Banach space of measurable functions $u : Z \to \mathbb{R}$ such that $\int_Z |u(z)|^p \, d\rho(z)$ is finite.

The *Nemitski* functional associated to W is defined as the map $I_0 : L^p(X, \nu) \to [0, +\infty[\cup \{+\infty\}$ given by

$$I_0[u] = \int_Z W(z, u(z)) d\rho(z). \tag{A.2}$$

Next proposition provides us with a straightforward method to study the subgradient (∂I_0). Let $q \in]1, +\infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$.

Proposition 12. *Assume that there is an element $u_0 \in L^p(Z, \nu)$ such that $\sup_{z \in Z} |u_0(z)| < +\infty$ and $I_0[u_0] < +\infty$. Given $u \in L^p(Z, \rho)$*

$$(\partial I_0)(u) = \{w \in L^q(Z, \rho) \mid \ w(z) \in (\partial W)(z, u(z)) \ \rho - \text{a.e.}\}. \tag{A.3}$$

Proof. See the proof of Prop. III.5.3 of [ET83b]. The proof is for Z interval of $\mathbb{R}$, but can be easily extended to arbitrary Z, compare with [ET74]. □

Bibliography

[ABDCBH97] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44:615–631, 1997.

[Arb95] A. Arbib, M. *The Handbook of Brain Theory and Neural Networks*. The MIT Press, Cambridge, MA, 1995.

[Aro50] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[BB98] M. Bertero and P. Boccacci. *Introduction to Inverse Problems in Imaging*. IOP Publishing, Bristol, 1998.

[BBL04a] O. Bousquet, S. Boucheron, and G. Lugosi. *Introduction to Statistical Learning Theory*, volume Lectures Notes in Artificial Intelligence 3176, pages 169–207. Springer, Heidelberg, Germany, 2004. A Wiley-Interscience Publication.

[BBL04b] O. Bousquet, S. Boucheron, and G. Lugosi. Theory of classification: A survey of recent advances. *to appear in ESAIM Probability and Statistics*, 2004.

[BBM99] A. Barron, L. Birge, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113:301–413, 1999.

[BC00] C. Burges and D. Crisp. Uniqueness of the SVM solution. In *Proceedings of the Twelfth Conference on Neural Information Processing Systems*, pages 223–229, Cambridge, MA, 2000. MIT Press.

[BC03] C. Burges and D. Crisp. Uniqueness theorems for kernel methods. *Neurocomputing*, 55:187–220, 2003.

[BDMP85] M. Bertero, C. De Mol, and E. R. Pike. Linear inverse problems with discrete data. I. General formulation and singular system analysis. *Inverse Problems*, 1(4):301–330, 1985.

[BDMP88] M. Bertero, C. De Mol, and E. R. Pike. Linear inverse problems with discrete data. II. Stability and regularisation. *Inverse Problems*, 4(3):573–594, 1988.

[BE02] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[BJM05] P.L. Bartlett, M.I. Jordan, and J.D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 2005. To appear. (Was Department of Statistics, U.C. Berkeley Technical Report number 638, 2003).

[BM02] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities. *Journal of Machine Learning Research*, 3:463–482, 2002.

[BMR06] T. Bissantz, N.and Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. Technical Report Preprint 2006-02, Institute for Mathematical Stochastics, University of Goettingen, 2006.

[Bou02] O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. 2002. PhD thesis, in press.

[BP05] F. Bauer and S. Pereverzev. Regularization without preliminary knowledge of smoothness and error behavior. *accepted in the European Journal of Applied Mathematics*, 2005.

[BPR05] F. Bauer, S. V. Pereverzev, and L. Rosasco. On regularization algorithms in leanring theory. Technical Report DISI-TR-05-18, DISI Universitá di Genova, december 2005. retrievable at http://www.disi.unige.it/person/RosascoL/.

[BS03] M. S. Birman and M. Solomyak. Double operators integrals in hilbert spaces. *Integr. Equ. Oper. Theory*, pages 131–168, 2003.

[BY02] P. Bühlmann and B. Yu. Boosting with the l_2-loss: Regression and classification. *Journal of American Statistical Association*, 98:324–340, 2002.

[CDV05a] A. Caponnetto and E. De Vito. Fast rates for regularized least-squares algorithm. Technical Report CBCL Paper 248/AI Memo 2005-033,,

Massachusetts Institute of Technology, Cambridge, MA, 2005. retrievable at http://cbcl.mit.edu/cbcl/publications/ai-publications/2005/AIM-2005-019.pdf.

[CDV05b] A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *submitted*, 2005.

[CDVT05] C. Carmeli, E. De Vito, and A. Toigo. Reproducing kernel hilbert spaces and mercer theorem. *eprint arXiv: math/0504071*, 2005. available at http://arxiv.org.

[CO90] D. Cox and F. O'Sullivan. Asymptotic analysis of penalized likelihood and related estimators. *Ann. Stat.*, 18:1676–1695, 1990.

[CR+05] A. Caponnetto, L. Rosasco, , F. Odone, and A. Verri. Support vectors algorithms as regularization networks. In *Proceedings of 13th European Symposium on Artificial Neural Networks*, 2005.

[CRDVV05] A. Caponnetto, L. Rosasco, E. De Vito, and A. Verri. Empirical effective dimension and optimal rates for regularized least squares algorithm. Technical Report CBCL Paper 252/AI Memo 2005-019,, Massachusetts Institute of Technology, Cambridge, MA, 2005. retrievable at http://cbcl.mit.edu/cbcl/publications/ai-publications/2005/AIM-2005-019.pdf.

[CS02a] F. Cucker and S. Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computationals Mathematics*, 2:413–428, 2002.

[CS02b] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

[CS05] A. Christmann and I. Steinwart. Consistency and robustness of kernel based regression. Technical Report Technical Report LA-UR-04-8797, Los Alamos National Laboratory, 2005. submitted for publication.

[CST00] N. Cristianini and J. Shawe Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.

[DKPT04] R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov. On mathematical methods of learning. Technical Report 2004:10, Industrial Mathematics Institute, Dept. of Mathematics University of South Carolina, 2004. retrievable at http://www.math/sc/edu/ imip/04papers/0410.ps.

[Dud02] R. M. Dudley. *Real Analysis and Probability*. Cambridge Uni. Press, Cambridge, 2002. 2nd edition.

[DVC05] E. De Vito and A. Caponnetto. Risk bounds for regularized least-squares algorithm with operator-valued kernels. Technical Report CBCL Paper 249/AI Memo 2005-015, Massachusetts Institute of Technology, Cambridge, MA, may 2005.

[DVCR05] E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5(1):59–85, February 2005.

[DVRC+04] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5:1363–1390, 2004.

[DVRC05a] E. De Vito, L. Rosasco, and A. Caponnetto. Discretization error analysis for tikhonov regularization. *to appear in Analysis and Applications*, 2005.

[DVRC+05b] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, May 2005.

[DVRCG04] E. De Vito, L. Rosasco, A. Caponnetto, and De Giovannini. Learning, regularization and ill-posed inverse problems. In *Proceedings of the Eighteenth Conference on Neural Information Processing Systems*, Cambridge, MA, 2004. MIT Press.

[EHN96] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[EPP00] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Adv. Comp. Math.*, 13:1–50, 2000.

[ET74] I. Ekeland and R. Teman. *Analyse convexe et problémes variationnels*. Gauthier-Villards, Paris, 1974.

[ET83a] I. Ekeland and T. Turnbull. *Infinite-dimensional Optimization and Convexity*. Chicago Lectures in Mathematics. The University of Chicago Press, Chicago, 1983.

[ET83b] I. Ekeland and T. Turnbull. *Infinite-dimensional Optimization and Convexity*. Chicago Lectures in Mathematics. The University of Chicago Press, Chicago, 1983.

[Fin99] L. Fine, T. *Feedforward Neural Network Methodology*. Springer-Verlag, 1999.

[FM01] G. Fung and O. L. Mangasarian. Proximal support vector machine classifiers. Technical Report 01-02, Data Mining Institute - University of Wisconsin - Madison, February 2001.

[FS97] Y. Freund and R.E. Schapire. A decision theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Science*, 1(55):119–139, 1997.

[Gir98] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

[GJP95] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269, 1995.

[GKW96] M. Györfi, L.and Kohler, A. Krzyzak, and H. Walk. *A Distribution-free Theory of Non-parametric Regression*. Springer Series in Statistics, New York, 1996, 1996.

[GP03] I. Guyon and Eliseeff M. P. An introduction to variable and feature selection. *Journal of Machine Learning Research*, Special Issue on Variable and Feature Selection(3):1157–1182, 2003.

[Gro84] C. W. Groetsch. *The theory of Tikhonov regularization for Fredholm equations of the first kind*, volume 105 of *Research Notes in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1984.

[Had02] J. Hadamard. *Bull. Univ. Princeton*, 13(49), 1902.

[Had23] J. Hadamard. Lectures on cauchy's problem in linear partial differential equations. *New Haven: Yale Univ. Press*, 1923.

[Han00] F. Hansen. Operator inequalities associated to jensen's inequality. *survey of "Classical Inequalities"*, pages 67–i98, 2000.

[HTF01] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[Kec01] V. Kecman. *Learning and Soft Computing*. The MIT Press, Cambridge, MA, 2001.

[Kur04] V. Kurkova. Learning from data as an inverse problem. In J. Antoch, editor, *COMPSTAT2004*, pages 1377–1384. Springer-Verlag, 2004.

[KW70] G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Stat.*, 41:495–502, 1970.

[Lan93] S. Lang. *Real and Functional Analysis*. Springer, New York, 1993.

[Lep90] O. Lepskii. A problem of adaptive estimation in guassian white noise. *Theory Probab. Appl.*, 36:454–466, 1990.

[LL04] J.M. Loubes and C. Ludena. Model selection for non linear inverse problems. *submitted to Probability Theory and Related Fields*, 2004.

[LLW02] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46:191–202, 2002.

[LM01] Y. J. Lee and O. L. Mangasarian. Ssvm: A smooth support vector machine for classification. *Computational Optimization and Applications*, 20(1):5–22, october 2001.

[MBBF00] L. Mason, J. Baxter, P. L. Bartlett, and M. Frean. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems 12*, pages 512–518, 2000.

[Men03] S. Mendelson. Estimating the performance of kernel classes. *Journal of Machine Learning Research*, 4:759–771, 2003.

[MM01] O. L. Mangasarian and D. R. Musicant. Data discrimination via nonlinear generalized support vector machines. In M. C. Ferris, O. L. Mangasarian, and J.S. Pang, editors, *Complementarity: Applications, Algorithms and Extensions*, pages 233–251. Kluwer Academic Publishers, 2001.

[MNPR04] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Statistical learning: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. Technical Report CBCL Paper 223, Massachusetts Institute of Technology, january revision 2004.

[MP02] P. Mathé and S. Pereverzev. Moduli of continuity for operator monotone functions. *Numerical Functional Analysis and Optimization*, 23:623–631, 2002.

[MP03] P. Mathé and S. Pereverzev. Geometry of linear ill-posed problems in variable hilbert scale. *Inverse Problems*, 19:789–803, June 2003.

[MP05a] P. Mathé and S. Pereverzev. Regularization of some linear ill-posed problems with discretized random noisy data. *accepted in Mathematics of Computation*, 2005.

[MP05b] C.A. Micchelli and M. Pontil. On learning vector–valued functions. *Neural Computation*, 17:177–204, 2005.

[MR03] R. Meir and G. Rätsch. *An Introduction to Boosting and Leveraging*, volume Lectures Notes in Artificial Intelligence 3176, pages 119–184. Springer, 2003.

[MRP02] S. Mukherjee, R. Rifkin, and T. Poggio. Regression and classification with regularization. *Lectures Notes in Statistics: Nonlinear Estimation and Classification, Proceedings from MSRI Workshop*, 171:107–124, 2002.

[NG99] P. Niyogi and F. Girosi. Generalization bounds for function approximation from scattered noisy data. *Adv. Comput. Math.*, 10:51–80, 1999.

[OC04] C.S. Ong and S. Canu. Regularization by early stopping. Technical report, Computer Sciences Laboratory, RSISE, ANU, 2004.

[PG92] T. Poggio and F. Girosi. A theory of networks for approximation and learning. In C. Lau, editor, *Foundation of Neural Networks*, pages 91–106. IEEE Press, Piscataway, N.J., 1992.

[PG00] S. Pontil, M.and Mukherjee and F. Girosi. On the noise model of support vector machine regression. In *Proc. of Algorithmic Learning Theory*, 2000.

[PMR+02] T. Poggio, S. Mukherjee, R. Rifkin, A. Rakhlin, and A. Verri. B. In J. Winkler and M. Niranjan, editors, *Uncertainty in Geometric Computations*, pages 131–141. Kluwer Academic Publishers, 2002.

[PRMN04] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in le arning theory. *Nature*, 428:419–422, 2004.

[PS85] I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for probabilities of large deviations. *Theory Probab. Appl.*, 30(1):143–148, 1985.

[RDVC+04] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural Computation*, 15(5):1063–1076, 2004.

[RDVV05] L. Rosasco, E. De Vito, and A. Verri. Spectral methods for regularization in learning theory. Technical Report DISI-TR-05-18, DISI, Universitá degli Studi di Genova, Italy, 2005. retrievable at http://www.disi.unige.it/person/RosascoL.

[RMP05] A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 3:397–419, 2005.

[Roc70] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, N.J., 1970.

[Rud91] W. Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. Mc Graw Hill, Princeton, 1991.

[Sch64] L. Schwartz. Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *J. Analyse Math.*, 13:115–256, 1964.

[SHS01] B. Schölkopf, R. Herbrich, and A.J. Smola. A generalized representer theorem. In D. Helmbold and B. Williamson, editors, *Neural Networks and Computational Learning Theory*, number 81, pages 416–426. Springer, Berlin, Germany, 2001.

[SP03] E. Schock and S. V. Pereverzev. On the adaptive selection of the parameter in regularization of ill-posed problems. Technical report, University of Kaiserslautern, august 2003.

[SPST+01] B. Schlkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[SS02] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[Ste03] I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.

[Ste04] I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *accepted on IEEE Transaction on Information Theory*, 2004.

[Ste05] I. Steinwart. How to compare different loss functions and their risks. Technical Report Technical Report LA-UR-05-7016, Los Alamos National Laboratory, 2005. submitted for publication.

[SVGDB+02] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.

[SZ03] S. Smale and D.X. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(1):1–25, 2003.

[SZ04] S. Smale and D.X. Zhou. Shannon sampling II : Connections to learning theory. *preprint*, 2004.

[SZ05] S. Smale and D.X. Zhou. Learning theory estimates via integral operators and their approximations. *submitted*, 2005. retrievable at http://www.tti-c.org/smale.html.

[TA77] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill Posed Problems*. W. H. Winston, Washington, D.C., 1977.

[TD99] D.M.J. Tax and R.P.W. Duin. Data domain description using support vectors. In *Proceedings of European Symposium on Artificial Neural Networks '99, Brugge*, 1999.

[TGSY95] A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov, and A. G. Yagola. *Numerical methods for the solution of ill-posed problems*, volume 328 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1995. Translated from the 1990 Russian original by R. A. M. Hoksbergen and revised by the authors.

[Tsy04] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

[Val84] L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vap98] V. N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

[vdG00] S. A. van de Geer. *Empirical Process in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2000.

[vdVW96] S. W. van de Vaart and J. A. Wellner. *Weak Convergence and Empirical Process Theory*. Springer Series in Statistics, New York, 1996. Springer-Verlag, New York, 1996.

[Wah90] G. Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[Wah98] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and randomized GACV. Technical Report 984, Department of Statistics, University of Wisconsin, 1998.

[WYZ04] Q. Wu, Y. Ying, and D.X. Zhou. Learning theory: from regression to classification. *preprint*, 2004. submitted for publication.

[YRC05] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *to be publiseh in Constructive Approximation*, 2005. retrievable at http://math.berkeley.edu/ yao/research.html.

[Zha01] T. Zhang. Convergence of large margin separable linear classification. In T.G. Len, T.K. Dietterich and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 357–363. MIT Press, 2001.

[Zha05] T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17:2077–2098, 2005.

[Zho03] D.X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Trans. Inform. Theory*, 49:1743–1752, 2003.

[ZY03] T. Zhang and B. Yu. Boosting with early stopping: convergence and consistency. Technical Report Technical Report 635, Department of Statistics, University of California at Berkeley, 2003.