Dual regularized total least squares in learning theory

Moussa Salman, S.

Award date:
2009

Dual Regularized Total Least Squares in Learning Theory

Sanaa Moussa Salman

January, 2009
Linz, Austria.

For my family: my parents, my brother and my sisters

Abstract

The problem in supervised learning or learning from example is to find a rule which allows to predict an output from a new input in the situation when a sample of input-output pairs is given. The problem of learning, as it was shown recently, can be reduced to a linear inverse problem which can be solved by regularization techniques. The aim of the master thesis is to apply some newly developed regularization technique called dual regularized total least squares (dual RTLS) in Learning Theory. A motivation to look for a new technique is that in some situations the prediction based on the classical methods, such as the Tikhonov regularization, is not so satisfactory. The numerical experiments show that the use of the dual RTLS technique is very efficient.

Acknowledgments

I wish to express my thanks to my supervisor, Prof. Dr. Sergei Pereverzyev. This thesis would not have been complete without his expert advice and unfailing patience.

A special thanks to Dr. Sergiy Pereverzyev jun. for his help and support throughout working on this thesis and for his patiently assisting.

I also thank all professors and teachers in Technical University of Eindhoven (TU/e) and Johannes Kepler University of Linz (JKU) for providing all the necessary knowledge during my studies.

I would like to express a special word of thanks to my parents, Moussa and Ne'ma, brother Ahmed and sisters, Hoda, Manal and Hend for their never ending moral support and prayers which always acted as a catalyst in my whole life.

Meanwhile, I would like to thank my friends in my home country and in Europe for their good accompany and sharing all moments with me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In supervised learning or learning from examples a machine is trained, instead of programmed, to perform a given task on a number of input-output pairs. So, the problem is that of finding a deterministic rule allowing to correctly predict the output when a new input is given. In a probabilistic setting, a fundamental problem, studied by Statistical Learning Theory, is how the chosen function estimates the output for new inputs .It was recently shown that learning from examples can be seen as the problem of solving a linear inverse problem from a finite dimensional discretization. What makes learning peculiar is that the discretization is stochastic and cannot be controlled. In fact in this context we demand the regularization algorithm to take care of the random discretization. It is known that Tikhonov regularization can be effectively used in the context of learning and many standard results in inverse problem can be easily carried over with minor modifications.

The main goal of learning from examples is to infer an estimator, given a finite sample of data drawn according to a fixed but unknown probabilistic input-output relation. The desired property of the selected estimator is to perform well on new data, that is, it should generalize. The fundamental work of Vapnik [2] shows that the key to obtain a meaningful solution to the above problem is to control the complexity of the solution space. Interestingly, this is the idea underlying regularization techniques for ill-posed inverse problems. A careful analysis shows a rigorous connection between learning and regularization for inverse problems. In this research we also use this connection.

Our contribution to the analysis of learning problems consists in the application of recently proposed regularization technique, that was, never used before in Learning Theory. In the dissertation we introduce this new technique in the context of learning and show how it performs in several situations of interest.

After recalling the connection between learning theory and inverse problems, we show that regularization techniques other than Tikhonov regularization, namely Dual Regularized Total Least Squares [5], [8], can be used in learning.

In order to make a prediction, we need only to know the so-called training set which is nothing but a collection of input-output pairs, denoted by $D_n = \{(x_i, y_i)\}_{i=1}^n$ and given by the system under study.

In a recent research [7], Tikhonov regularization was employed to perform the prediction, however, the results was not satisfactory in some situations. More precisely, the quality of the prediction for inputs being beyond the scope of a training set is rather poor [see Figure 5.3].

The main thrust of this work is to apply the dual regularized total least squares (dual RTLS) to construct the approximation in a form of the neural network

$$f(x) = \sum_{i=1}^{n} c_i K(x; x_i),$$

where K is a reproducing kernel generating a network.

In more details, the thesis is organized as follows. In Chapter 2 we discuss the relation between the Learning theory and regularization techniques. It turns out that a learning problem can be reduced to an ill-posed linear operator equation in some small space, namely reproducing kernel Hilbert space (RKHS). Thus, some properties of the RKHS will be discussed.

As a matter of fact, the problem of approximating a function from sparse data is ill-posed. Thus, a short introduction about ill-posed problem will be given in Chapter 3 together with the discussion of the proposed regularization method for our research, namely the dual RTLS.

In addition, a good choice of the regularization parameters is crucial to assure a good approximation. Thus, in Chapter 4 we will construct a selection algorithm which is able to determine the optimal parameters.

The performance of our algorithm will be illustrated in Chapter 5, where we

use the examples from [3], which appear to be problematic for a treatment with standard regularization technique, such as Tikhonov regularization [7].

Chapter 2

The problem of learning as an ill-posed linear operator equation

This chapter reviews that a learning problem can be reduced to an ill-posed linear operator equation which can be solved by regularization techniques.

2.1 The problem of learning

In recent years, there has been an increasing interest in learning theory. This was a logical result of the so-called greatest problem of science today which refers to the problem of understanding intelligence. In [4] it is written:

The problem of learning represents a gateway to understanding intelligence in brains and machines, to discovering how the human brain works, and to making intelligent machines that learn from experience and improve their competence as children do.

Learning from examples, or supervised leaning refers to systems that are trained instead of programmed with a set of examples which consists of input-output pairs. Training means choosing a function which best describes the relation between the inputs and the outputs.

Many applications of systems that could learn from examples can be found. For example , a car manufacturer may want to have in its models a system to detect pedestrians about to cross the road to alert drivers to a possible danger while driving in downtown traffic. Such a system could be trained with positive and

negative examples: images of pedestrians and images without pedestrians. Actually, software trained in this way has been tested in an excremental car. It runs on a PC in the trunk and looks at the road in front of the car through a digital camera [4].

What we assume in the above example is a machine that is trained instead of programmed to perform a task. The only thing which is available here is the data of the form $(x_i, y_i)_{i=1}^n$. To be more specific, what we mean by training is synthesizing a function that represents the relation between the inputs x_i and the corresponding outputs y_i. One question that needs to be asked, however, is how well this function estimate the outputs for previously unseen inputs? In other words, a challenging problem within machine learning is how to make good inferences from data sets in which pieces of information are missing.

The main goal of machine learning is to develop general algorithms of practical value. Such algorithms should be efficient. In addition, Learning algorithms should also be as general purpose as possible. So, we are looking for algorithms that can be easily applied to a broad class of learning problems. Of primary importance, we want the result of learning to be a prediction rule that is as accurate as possible in the predictions that it makes. In other words, we want the computer to find prediction rules that are easily understandable by human experts.

As mentioned , machine learning can be thought of as programming by example. The central question which arises here: What is the advantage of machine learning over direct programming? There are two main reasons

• Firstly, the results of using machine learning are often more accurate than what can be created through direct programming. This is because machine learning algorithms are data driven, and are able to examine large amounts of data.

• Secondly, a human expert is likely to be guided by imprecise impressions or perhaps an examination of only a relatively small number of examples. In Figure2.1 a diagram of a typical learning problem is shown.
In the next section we will see that probability theory plays a key role in learning theory.

2.2 Statistical learning theory

The main goal of statistical learning theory is to provide a framework for studying the problem of inference, that is of gaining knowledge, making predictions,

making decisions or constructing models from a set of data. This is studied in a statistical framework in which there are assumptions of statistical nature about the underlying phenomena (in the way the data is generated).

Assume we have two sets of variables $\mathbf{x} \in X \subseteq \mathbb{R}^d$ and $y \in Y \subseteq \mathbb{R}$ which are related by a probabilistic relationship. The relation is called probabilistic because in general, an element of X does not determine uniquely an element of Y, but rather a probability distribution on Y. In order to formalize this we introduce a probability distribution $P(x, y)$ which is defined over the set $X \times Y$. Unfortunately, this probability distribution is unknown and under very general condition it can be written as $P(x, y) = P(x)P(y|x)$ where $P(x)$ is the marginal probability of x and $P(y|x)$ is the conditional probability of y given x.

As just mentioned, the probability distribution P is unknown, however, examples of the probabilistic relationship are provided. In fact, what we know is a data set $D_n = (x_i, y_i)_{i=1}^n$, called also the training data, which is obtained by sampling n times the set $X \times Y$ according to $P(x, y)$. In other words, the data set $D_n = (x_i, y_i)$, $i = 1, \ldots, n$, drawn i.i.d. according to unknown probability distribution P on $X \times Y$.

The problem of learning can be analyzed in two basic steps as follows

1. Given the data set $D_n \equiv \{(x_i, y_i)_{i=1}^n \in X \times Y\}$.

2. Providing an estimator, that is, a function $f : X \to Y$, that can be used in the sense that given any value of $x \in X$, a value of y can be predicted.

To solve the learning problem in view of the statistical learning theory, a risk functional should be defined. The **latter** measures the average amount of error associated with an estimator and then to look for the estimator among the allowed ones with the lowest risk. the average error is given as

$$I[f] \equiv \int_{X,Y} V(y, f(x))P(x, y)\, dx\, dy.$$

and is called expected risk, where $V(y, f(x))$ is the loss function. A natural choice for the loss function is the squared loss function $V(y, f(x)) = (f(x) - y)^2$. So, the expected risk can be written as

$$I[f] \equiv \int_{X,Y} (f(x) - y)^2 P(x, y)\, dx\, dy. \tag{2.1}$$

The expected risk is assumed to be defined on $L^2(X, P(x)dx)$. The function which minimizes the expected risk in $L^2(X, P(x)dx)$ is denoted by f_0 and is given by

$$f_0(x) = arg \min_{L^2(X,P(x)dx)} I[f].$$

It is important to mention that the function f_0 is our ideal estimator which is often called the target function. The key problem here is that this function can not be found in practice since the probability distribution $P(x, y)$ is unknown. The only information available is a sample of the target function, that is, the data set D_n. To overcome this problem we need an induction principle that used to learn from the limited number of the training data we have.

Vapnik developed statistical learning theory built on the so-called empirical risk minimization (ERM) induction principle [9] which consists in using the data set D_n to build a stochastic approximation of the expected risk,usually called the empirical risk, defined as

$$I_{emp}[f; n] = \frac{1}{n} \sum_{i=1}^{n} V(y_i, f(x_i)). \tag{2.2}$$

The central question of the theory is whether the expected risk of the minimizer of the empirical risk in $L^2(X, P(x)dx)$ is close to the expected risk of f_0. The theory finds under what conditions the method of ERM satisfies

$$\lim_{n\to\infty} I_{emp}[\widehat{f}_n; n] = \lim_{n\to\infty} I[\widehat{f}_n] = I[f_0] \tag{2.3}$$

in probability, where $\widehat{f}_n$ denotes the minimizer of the empirical risk (2.2) in $L^2(X, P(x)dx)$.

A necessary and sufficient condition for the limits in (2.3) to hold true in probability is

$$\lim_{n\to\infty} P\{\sup_{f\in L^2(X,P(x)dx)} (I[f] - I_{emp}[f; n]) > \varepsilon\} = 0 \ \ \forall \varepsilon > 0.$$

This condition is known as one-sided uniform convergence in probability of empirical risk to expected risk in $L^2(X, P(x)dx)$. Typically in literature the two-sided uniform convergence in probability:

$$\lim_{n\to\infty} P\{\sup_{f\in L^2(X,P(x)dx)} |I[f] - I_{emp}[f; n]| > \epsilon\} = 0 \quad \forall \epsilon > 0, \tag{2.4}$$

is considered. If $L^2(X, P(x)dx)$ is very large, we can always find $\widehat{f}_n \in L^2(X, P(x)dx)$ with zero empirical error. Nevertheless, this does not guarantee that the expected

risk of $\widehat{f_n}$ is also close to zero, or close to $I[f_0]$.

There is an unambiguous relationship between empirical risk $I_{emp}[f;n]$ and expected risk $I[f]$ which was first discussed by Vapnik and Chervonenkis with the help of the VC-dimension [9].

Definition 2.1. *The VC-dimension of a set* $\{\theta(f(x)), f \in L^2(X, P(x)dx), x \in X\}$*, of indicator functions is the maximum number* h *of vectors* $\boldsymbol{x}_1, ..., \boldsymbol{x}_h \in X$ *that can be separated into two classes in all* 2^h *possible ways using functions of the set, where* $\theta(.)$ *is the Heaviside function .*

The VC-dimension was first defined for the case of indicator functions and was then extended to real valued functions. If, for any number N, it is possible to find N points $\mathbf{x}_1, ..., \mathbf{x}_N$ that can be separated in all 2^N possible ways, we say that the VC-dimension is infinite.

Definition 2.2. *Let* $A \leq V(y, f(x)) \leq B$*,* $f \in L^2(X, P(x)dx)$*, with* A *and* $B < \infty$*. The VC-dimension of the set* $\{V(y, f(x)), f \in L^2(X, P(x)dx)\}$ *is defined as the VC-dimension of the set of the indicator functions* $\{\theta(V(y, f(x)) - \alpha), \alpha \in (A, B)\}$*.*

Vapnik and Chervonenkis studied the relation between the empirical risk and expected risk in a hypothesis space [9]. Actually, they suggest a method which does not only minimizes the empirical risk but also minimizes the complexity of the hypothesis space. This method is called structural risk minimization (SRM). The idea of SRM is to define a nested sequence of hypothesis spaces $H_1 \subset H_2 \subset ... \subset H_{m(n)}$ with $m(n)$ a non-decreasing integer function of n, where each hypothesis space H_i has finite VC-dimension and larger than that of all previous sets. Thus, if h_i is the VC-dimension of the space H_i , then $h_1 \leq h_2 \leq ... \leq h_{m(n)}$. For each hypothesis space H_i the solution of the learning problem is

$$\hat{f}_{i,n} = \min_{f \in H_i} I_{emp}[f;n].$$

Then in SRM one is looking for an appropriate choice of $m(n)$, so that as $n \to \infty$ and $m(n) \to \infty$, the expected risk of the solution of the method approaches in probability the minimum of the empirical risk. However, in practice one usually [2] uses as hypothesis space sets of bounded functions such that $H_i = H_{A_i} = \{f : ||f|| < A_i\}$, $i = 1, 2, \ldots, m$, where $||.||$ is some appropriate norm. Thus, in order

to use the standard SRM method we need to know the VC-dimension of such space under corresponding loss functions. Unfortunately, it can be shown that when the loss function V is $(y - f(x))^2$, the VC-dimension of $V(y, f(x))$ with f in $H_A = \{f : ||f|| < A\}$ does not depend on A, and is infinite if the corresponding space equipped with the norm $||.||$ is infinite dimensional.

Thus, it is impossible to use the SRM with this kind of hypothesis spaces: in the case of finite dimensional spaces, the norms of f can not be usually used to define a structure of spaces with different VC-dimensions. So, in many practical applications SRM cannot be used directly.

The way out of this situation is related with some other view on the learning problems, when it is viewed as a linear ill-posed operator equation. This approach is outlined below.

2.3 Background of regularization theory

2.3.1 Ideal predictor

To derive the algorithm corresponding to learning problem we start with the following important observation.

Proposition 2.1. *The expected risk which is given by (2.1) can be rewritten as*

$$I[f] = ||f - f_0||^2_{L^2(X,P(x)dx)} + I[f_0],$$

for any $f \in L^2(X, P(x)dx)$.

Proof. Since we have

$$\begin{aligned} I[f] &= \int_X \int_Y (f(x) - y)^2 P(x, y)\, dy\, dx \\ &= \int_X \int_Y (f(x) - f_0(x) + f_0(x) - y)^2 P(x, y)\, dy\, dx \\ &= \int_X (f(x) - f_0(x))^2 P(x)\, dx + \int_X \int_Y (f_0(x) - y)^2 P(x, y)\, dy\, dx \\ &+ 2 \int_X \int_Y (f(x) - f_0(x))(f_0(x) - y) P(y|x) P(x)\, dy\, dx \end{aligned}$$

where we use the relation

$$P(x, y) = P(x)P(y|x).$$

Now if we look at the term $2\int_X \int_Y (f(x) - f_0(x))(f_0(x) - y)P(y|x)P(x)\, dy\, dx$, it is not difficult to show that it vanishes:

$$\int_Y (f(x) - f_0(x))(f_0(x) - y)P(y|x)\, dy$$

$$\begin{aligned} &= (f(x) - f_0(x))(f_0(x) - \int_Y yP(y|x)\, dy) \\ &= 0. \end{aligned}$$

Hence,

$$\begin{aligned} I[f] &= \int_X (f(x) - f_0(x))^2 P(x)\, dx + \int_X \int_Y (f_0(x) - y)^2 P(x, y)\, dy\, dx \\ &= ||f - f_0||^2_{L^2(X,P(x)dx)} + I[f_0]. \end{aligned}$$

□

Our idea now is to look for the minimizer of (2.1) in a space which is smaller than $L^2(X, P(x)dx)$. For this, we define a Hilbert subspace $\mathcal{H} \subset L^2(X, P(x)dx)$ which is specified in our research to be a Reproducing Kernel Hilbert Space (RKHS). Introduce the inclusion operator $\mathfrak{I} : \mathcal{H} \to L^2(X, P(x)dx)$ and its adjoint $\mathfrak{I}^* : L^2(X, P(x)dx) \to \mathcal{H}$. For any $f \in \mathcal{H}$, it is possible to write $||f - f_0||_{L^2(X,P(x)dx)}$ as $||\mathfrak{I}f - f_0||_{L^2(X,P(x)dx)}$. Now from the least square form we know that the minimizer $f_{\mathcal{H}}$ of the expected risk $I[f]$ over $\mathcal{H}$ solves the equation

$$\mathfrak{I}^*\mathfrak{I}f = \mathfrak{I}^* f_0. \tag{2.5}$$

Equation (2.5) shows the possibility of approximating $f_{\mathcal{H}}$ from f_0 by regularization techniques. unfortunately, the target function f_0 is unknown, only its discrete version is available, namely the data set. To this end, we discretize the above equation using the data set and then apply regularization techniques to solve it.

To be able to use the data set D_n for the discretization of (2.5) in $\mathcal{H}$, the latter one should have some special structure. Normally, a function evaluation $f(x_i)$ should be considered as a linear bounded functional in $\mathcal{H}$. The spaces with such a structure are called Reproducing Kernel Hilbert Spaces (RKHS).

2.3.2 Reproducing Kernel Hilbert Spaces

Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ is a Hilbert space of functions defined over some bounded domain $X \subset \mathbb{R}$ with the property that, for each $x \in X$, the evaluation functionals $\mathcal{F}_x$ defined as

$$\mathcal{F}_x[f] = f(x) \qquad \forall f \in \mathcal{H}$$

are linear, bounded functionals. The boundedness means that there exists a $U \in \mathbb{R}^+$ such that

$$|\mathcal{F}_x[f]| = |f(x)| \leq U \parallel f \parallel_{\mathcal{H}},$$

for all f in the RKHS.

To every RKHS $\mathcal{H}$ there corresponds a unique positive definite function $K(x, y)$ of two variables in X, called the reproducing kernel of $\mathcal{H}$, that has the following reproducing property:

$$f(x) = \langle f(y), K(x, y)\rangle_{\mathcal{H}} \ \ \forall f \in \mathcal{H}, \tag{2.6}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the scalar product in $\mathcal{H}$ [10].

Assume that we have a sequence of positive numbers λ_n and linearly independent functions $\phi_n(x)$ such that the function $K(x, y)$ admits the representation

$$K(x, y) = \sum_{n=0}^{\infty} \lambda_n \phi_n(x)\phi_n(y), \tag{2.7}$$

where the series is uniformly convergent.

In view of (2.7), RKHS can be seen as the set of functions of the form

$$f(x) = \sum_{n=0}^{\infty} a_n \phi_n(x),$$

for $a_n \in \mathbb{R}$, and define the scalar product in our space to be

$$\langle \sum_{n=0}^{\infty} a_n \phi_n(x), \sum_{n=0}^{\infty} d_n \phi_n(x)\rangle_{\mathcal{H}} \equiv \sum_{n=0}^{\infty} \frac{a_n d_n}{\lambda_n}. \tag{2.8}$$

In fact we have

$$\langle f(y), K(x, y)\rangle_{\mathcal{H}} = \sum_{n=0}^{\infty} \frac{a_n \lambda_n \phi_n(x)}{\lambda_n} = \sum_{n=0}^{\infty} a_n \phi_n(x) = f(x),$$

hence equation (2.6) is satisfied.

It is easy to show that whenever we have a function K of the form (2.7), it is possible to construct a RKHS as shown above. Vice versa, for any RKHS there is a unique function K and corresponding λ_n, ϕ_n, that satisfy (2.7). Moreover, equation(4.9) shows that the norm of the RKHS has the form

$$||f||_K^2 = \sum_{n=0}^{\infty} \frac{a_n^2}{\lambda_n}.$$

In shorter words, the Hilbert space $L^2(X, P(x)dx)$ is too "big" for our purposes, containing too many non-smooth functions. One approach to obtaining restricted, smooth spaces is the Reproducing Kernel Hilbert Space (RKHS) approach. A RKHS is "smaller" than a general Hilbert space. It is a Hilbert space of point-wise defined functions which can be completely characterized by a symmetric positive definite function $K : X \times X \to \mathbb{R}$, namely the kernel.

2.3.3 Discretization of a linear operator

Redefine the inclusion operator $\mathfrak{I} = \mathfrak{I}_{\mathcal{H}_K} : \mathcal{H} \to L^2(X, P(x)dx)$, where $\mathcal{H}_K$ denotes the RKHS here. Now we want to discretize the equation

$$\mathfrak{I}^*_{\mathcal{H}_K} \mathfrak{I}_{\mathcal{H}_K} f = \mathfrak{I}^*_{\mathcal{H}_K} f_0, \tag{2.9}$$

Moreover, we define the covariance operator $T : \mathcal{H} \to \mathcal{H}$ such that $T = \mathfrak{I}^*_{\mathcal{H}_K} \mathfrak{I}_{\mathcal{H}_K}$. In addition, T can be written as [1]

$$T = \int_X \langle ., K_x \rangle_{\mathcal{H}} K_x P(x)\, dx.$$

The operator T can be proved to be positive trace class operator and hence compact. Define the sampling operator $S_x : \mathcal{H} \to \mathbb{R}^n$ by $(S_x f)_i = f(x_i) = \langle f, K_{x_i} \rangle_{\mathcal{H}}$; $i = 1, ..., n$. Moreover, we define the adjoint operator $S_x^* : \mathbb{R}^n \to \mathcal{H}$, and the operator $T_x : \mathcal{H} \to \mathcal{H}$ such that $T_x = S_x^* S_x$. It follows that for $\mathbf{y} = (y_1, ..., y_n)$

$$S_x^* \mathbf{y} = \frac{1}{n} \sum_{i=1}^{n} K_{x_i} y_i, \;\; T_x = \frac{1}{n} \sum_{i=1}^{n} \langle ., K_{x_i} \rangle_{\mathcal{H}} K_{x_i}. \tag{2.10}$$

We discretize equation (2.9) using training set, and then apply regularization techniques to solve it. Actually, S_x is nothing else but a discrete analog of the mapping $\mathfrak{I}_K$, that is , if we replace $\mathfrak{I}_{\mathcal{H}_K}$ by S_x , we replace the target function f_0 by the discrete data $\mathbf{y}$, that is, we get

$$\mathfrak{I}_{\mathcal{H}_K} f = f_0 \Rightarrow S_x f = \mathbf{y}.$$

In a similar way, the continuous operator $T = \mathfrak{I}^*_{\mathcal{H}_K} \mathfrak{I}_{\mathcal{H}_K}$ can be replaced by the discrete operator T_x:

$$Tf = \mathfrak{I}^*_{\mathcal{H}_K} f_0 \Rightarrow T_x f = S^*_x \mathbf{y}.$$

So, with the discretized equation $T_x f = S^*_x \mathbf{y}$ it is not difficult to approach the target function with the help of regularization methods. In our research the dual Regularized Total Least Squares method (dual RTLS) is employed. Tikhonov regularization method was employed in many researches but results were not so satisfactory in some situations as we will see later. That is why we look for a better method in the current research to get better approximations . In the next chapter we introduce the dual RTLS method in some details.

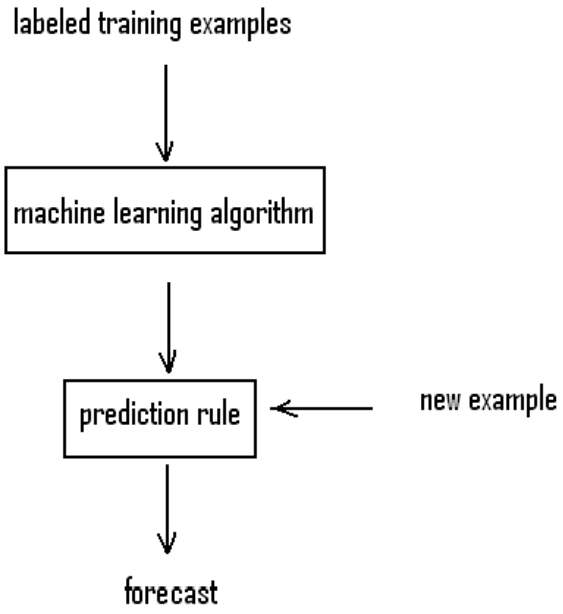

Figure 2.1: Diagram of a typical learning problem.

Chapter 3

Regularization techniques

As mentioned before, regularization techniques can be used to solve learning problems. Several studies investigating regularization have been carried out on RKHS. In this chapter we study the possibility of solving ill-posed problems with noisy right-hand side and a noisy operator. We will discuss the proposed **regularization** method for our research, namely the dual Regularized Total Least Squares (dual RTLS).

Since as mentioned in the previous chapter, the problem of approximating a function from sparse data is ill-posed, we discuss in the first section the ill-posed problems.

3.1 Ill-posed problems

Ill-posed problems arise in many context and have important applications in science and engineering. We consider ill-posed problems having the form of a linear operator equation

$$Ax = y, \tag{3.1}$$

where A is a compact linear operator between Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$.

Ill-posed problems are mathematical problems which do not satisfy Hadamard's definition of well-posedness:

- $\mathcal{R}(A) = \mathcal{Y}$ For all admissible data, a solution exsits.
- $\mathcal{N}(A) = 0$ For all admissible data, the solution is unique.
- $A^{-1} \in \mathcal{L}(\mathcal{Y}, \mathcal{X})$ The solution depends continuously on the data.

If one wants to approximate a problem whose solution does not depend continuously on the data by a traditional numerical method as one would use for well-posed problems, it is expected that the numerical method becomes unstable. A (partial) remedy for this is the use of "regularization methods". In general terms, regularization is the approximation of an ill-posed problem by a family of neighboring well-posed problems.

Remark 3.1. *For a compact linear operator with non-closed range (e.g., for an integral operator with a non-degenerate L^2 kernel), the solution of $Ax = y$ does not depend continuously on the right-hand side; the equation is ill-posed.*

Definition 3.1. *Let $A : \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator.*

1. *$x \in \mathcal{X}$ is called least-squares solution of $Ax = y$ if*

$$||Ax - y|| = inf\{||Az - y|| \mid z \; in \; \mathcal{X}\}.$$

2. *$x \in \mathcal{X}$ is called best approximate solution of $Ax = y$ if x is a least squares solution of $Ax = y$ and*

$$||x|| = inf\{||z|| \mid z \text{ is least-squares solution of } Ax = y\}$$

holds.

Definition 3.2. *The Moore-Penrose generalized inverse $A^\dagger$ of $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ is defined as the unique linear extension of $\tilde{A}^{-1}$ to*

$$\mathcal{D}(A^\dagger) := \mathcal{R}(A) \dotplus \mathcal{R}(A)^\perp,$$

with

$$\mathcal{N}(A^\dagger) = \mathcal{R}(A)^\perp,$$

where

$$\tilde{A} := A|_{\mathcal{N}(A)^\perp} : \mathcal{N}(A)^\perp \to \mathcal{R}(A).$$

We want to approximate the best-approximate solution $x^\dagger := A^\dagger y$ of equation (3.1) for a specific right-hand side y in the situation that:

• The exact data y is not known precisely, but that only an approximation y^δ with

$$||y^\delta - y|| \leq \delta \tag{3.2}$$

is available, where y^δ is called the noisy data and δ the noise level.
• The exact operator A is not known, but we have some noisy operator A_h with

$$||A - A_h|| \leq h. \tag{3.3}$$

In other words, we are looking for some approximation, say x^δ, of the solution of (3.1) which depends continuously on the noisy data y^δ, so that it can be computed in a stable way. As a result, x^δ tends to the solution of (3.1), when δ tends to zero.

The range of the operator A, $\mathcal{R}(A)$, is assumed to be non-closed. Thus, the solution $x^\dagger$ of equation (3.1) does not depend continuously on the data. So, problem (3.1), (3.2), (3.3) requires the application of special regularization techniques for its numerical treatment. Below, we will discuss these regularization techniques in some details.

3.2 Least Squares

Least squares (LS) technique, also known as Ordinary Least Squares (OLS), is the simplified method used to solve systems of the form $A_h x = y^\delta$, under the constraints (3.2) and (3.3). Least squares is often applied in statistical contexts, particularly regression analysis. This technique can be interpreted as **a data fitting technique** . The best fit in the least-squares sense is that instance of the model for which the sum of squared residuals has its least value, where a residual is the difference between an observed value and the value given by the model.

In other words, it is a mathematical procedure for finding the best-fitting curve to a given set of points by minimizing the sum of the squares of the offsets "the residuals" of the points from the curve. The sum of the squares of the offsets is used instead of the offset absolute values because this allows the residuals to be treated as a continuous differentiable quantity.

The least-squares approximate solution of $Ax = y$ is given by

$$x_{ls} = (A^T A)^{-1} A^T y.$$

This is the unique $x \in \mathbb{R}^n$ that minimizes $||Ax - y||$.

The LS problem can also be seen as follows: We look for x, y such that $y = A_h x$:

$$\min_{x,y} ||y - y^\delta||_2 \quad \text{subject to} \quad y = A_h x.$$

Since squares of the offsets are used, outlying points can have a disproportionate effect on the fit, a property which may or may not be desirable depending on the problem at hand.

3.3 Tikhonov regularization

Tikhonov regularization is one of the most widely applied methods for solving ill-posed problems. In this method a regularized approximation $x_\alpha^{\delta,h}$ is obtained by solving the minimization problem

$$\min_{x \in X} J_\alpha(x) = ||A_h x - y^\delta||^2 + \alpha ||x||^2,$$

where $\alpha > 0$ is the regularization parameter to be chosen properly. Hence, in Tikhonov's method the regularized approximation is given by

$$x_\alpha^{\delta,h} = (A_h^* A_h + \alpha \mathbf{I})^{-1} A_h^* y^\delta.$$

where $\mathbf{I}$ is the identity operator.

3.4 Total Least Squares

Total Least Squares (TLS) is another method for treating a problem of the form of linear equations $Ax = y$, where both the operator A and the right-hand side y are contaminated by noise. In practical situations, the linear system is often ill-conditioned. For example, this happens when the system is obtained by discretization of ill-posed problems such as integral equations of the first kind.

The basic idea of the classical total least squares problem is to find some estimate $(\hat{x}, \hat{y}, \hat{A})$ for $(x^\dagger, y, A)$ using given data (y^δ, A_h). This is done by solving the constrained minimization problem [11], [12], [5].

$$||A - A_h||^2 + ||y - y^\delta||^2 \to \min \quad \text{subject to} \quad Ax = y. \tag{3.4}$$

As mentioned, in many practical applications, all data are contaminated by noise, which motivates the use of TLS. Efficient and reliable numerical methods to compute the TLS solution are based on the singular value decomposition (SVD).

Difficulties arise, however, because of the ill-posedness of equation(3.1) as it may happen that there does not **exists** any solution $\hat{x}$ of TLS problem(3.4) in the space $\mathcal{X}$. Moreover, if a solution exists, it may be far from the desired solution $x^{\dagger}$**.**

3.5 Regularized total least squares

For the reasons just mentioned, we restrict the set of admissible solutions by looking for approximations $\hat{x}$ that belong to some prescribed set K, which is the basic idea of regularized toltal least squares (RTLS).

The set K can be defined in several ways. In the simplest case it is a ball $K = \{x \in \mathcal{X} \mid ||Bx|| \leq R\}$ with prescribed radius R. This leads us to the RTLS problem in which, as we mentioned above, some estimate $(\hat{x}, \hat{y}, \hat{A})$ for $(x^{\dagger}, y, A)$ is determined by solving the constrained minimization problem

$$||A - A_h||^2 + ||y - y^{\delta}||^2 \rightarrow min \quad \text{subject to} \quad Ax = y, \quad ||Bx|| \leq R. \tag{3.5}$$

Special Case: When the operator A_h is exactly given, that is, $A_h = A$, the idea of this technique leads us to the method of quasi-solution of Ivanov (see [16]), where $\hat{x}$ is determined by solving the constrained minimization problem

$$||Ax - y^{\delta}||^2 \rightarrow \min \quad \text{subject to} \quad x \in K.$$

This approximation $\hat{x}$ is sometimes called K-constrained least squares solution.

Theorem 3.1. *[11], [13], [14], [5] If the constraint $||Bx||_2 \leq R$ of the RTLS problem (3.5) is active, then the RTLS solution $x = \hat{x}$ satisfies the equations*

$$(A_h^{\top} A_h + \alpha B^{\top} + \beta I)x = A_h^{\top} y^{\delta} \ \textit{and} \ ||Bx||_2 = R, \tag{3.6}$$

where the parameters α and β satisfy

$$\alpha = \mu(1 + ||x||_2^2) \ \textit{and} \ \beta = -\frac{||A_h x - y^{\delta}||_2^2}{1 + ||x||_2^2}, \tag{3.7}$$

where $\mu > 0$ is the Lagrange multiplier. Moreover,

$$\beta = \alpha R^2 - {y^{\delta}}^{\top}(y^{\delta} - A_h x) = -||A - A_h||_F^2 - ||y - y_{\delta}||_2^2, \tag{3.8}$$

where $||.||_F$ denotes Frobenius norm.

3.6 Dual regularized total least squares

A serious weakness of the RTLS is that it requires a reliable bound R for the norm $||Bx^{\dagger}||$. Unfortunately, such a bound is unknown in many practical applications. In different applications, however, reliable bounds for δ and h which appear in (3.2) and (3.3) are known. Thus, it makes sense to look for approximation $(\hat{x}, \hat{y}, \hat{A})$ which satisfy side conditions $Ax = y$, $||y^{\delta} - y|| \leq \delta$ and $||A - A_h|| \leq h$.

Choosing from the set of solutions the element which minimizes $||Bx||$ leads us to a problem in which some estimate $(\hat{x}, \hat{y}, \hat{A})$ for $(x^{\dagger}, y, A)$ is determined by solving the constrained minimization problem

$$||Bx|| \rightarrow min \quad \text{subject to} \quad Ax = y, ||y^{\delta} - y|| \leq \delta, ||A - A_h|| \leq h. \tag{3.9}$$

The method (3.9) can be seen as the dual of (3.5). Therefore, we call the later method as the dual regularized total least squares problem (dual RTLS problem). The dual regularized total least squares (dual RTLS) can be characterized as a special multi-parameter regularization method where one of the two regularization parameters is negative.

Theorem 3.2. *[5] If the two constraints $||y^{\delta} - y||_2 \leq \delta$ and $||A - A_h||_F \leq h$ of the dual RTLS problem (3.9) are active, then the dual RTLS solution $x = \hat{x}$ of the problem (3.9) is a solution of the equation*

$$(A_h^{\top} A_h + \alpha B^{\top} B + \beta I)x = A_h^{\top} y^{\delta}, \tag{3.10}$$

with parameters α and β solving the system

$$||A_h x^{\delta,h}(\alpha, \beta) - y^{\delta}|| = \delta + h||x^{\delta,h}(\alpha, \beta)||, \qquad \beta = -\frac{h(\delta + h||x^{\delta,h}(\alpha, \beta)||)}{||x^{\delta,h}(\alpha, \beta)||}, \tag{3.11}$$

where $x^{\delta,h}(\alpha, \beta)$ is the solution of (3.10) for fixed α, β.

As we already have noticed, both of the RTLS problem and its dual need one more regularization parameter than in Tikhonov. In our research we will restrict ourselves to the dual RTLS problem (3.9). In the next chapter we will discuss the computational aspects of the dual RTLS.

Chapter 4

Selection of the regularization parameters

The purpose of this chapter is to discuss the dual RTLS from a computational point of view. So, we present a strategy for selecting the two regularization parameters α and β which were introduced in the previous chapter. More precisely, we **describe** a model function of two variables for the dual RTLS as it is discussed in [6].

It can be understood from Theorem 3.2 that a realization of the dual RTLS **involves** with solving a highly nonlinear system of equations (3.11) [6]. It is observed that the first equation of (3.11) is similar to one appearing in the discrepancy principle for determining a regularization parameter in one parameter regularization methods applied to equations with noisy operator, where a model function approach has been proposed [17]. In the next section we will derive an appropriate form of a model function of two variables and see how can we use it for solving the system (3.11).

4.1 A model function method

Assume that a domain $\sum \subset \mathbb{R}^2$ is given such that (3.10) has a unique solution $x = x^{\delta,h}(\alpha, \beta)$ for any $(\alpha, \beta) \in \sum$. This solution is continuously differentiable with respect to both α and β. We know already from Theorem 3.2 that if the constraints are active, the dual RTLS solution $\hat{x} = x^{\delta,h}(\alpha, \beta)$ of problem(3.9) can be obtained by solving the minimization problem

$$\min_{x \in X} J_{\alpha,\beta}(x), \qquad J_{\alpha,\beta}(x) = ||A_h x - y^\delta||^2 + \alpha||Bx||^2 + \beta||x||^2,$$

with regularization parameters (α, β) chosen by the following a posteriori rule:
Dual RTLS rule: Choose (α, β) by solving the system (3.11).
It is clear that the Dual RTLS rule is a special multi-parameter choice rule of a posteriori type for choosing both regularization parameters α and β in Tikhonov's functional $J_{\alpha,\beta}$. For fixed $\alpha, \beta \in \sum$ the solution $x^{\delta,h}(\alpha, \beta)$ of the minimization problem $J_{\alpha,\beta}(x) \to \min$ is equivalent to the solution of the regularized equation (3.10), or equivalent to the solution of the variational equation

$$\langle A_h x, A_h g\rangle + \alpha\langle Bx, Bg\rangle + \beta\langle x, g\rangle = \langle y^\delta, A_h g\rangle \quad \forall g \in X. \tag{4.1}$$

Substituting $x^{\delta,h}(\alpha, \beta)$ into $J_{\alpha,\beta}(x)$ implies the following cost function

$$F(\alpha, \beta) = ||A_h x^{\delta,h}(\alpha, \beta) - y^\delta||^2 + \alpha||Bx^{\delta,h}(\alpha, \beta)||^2 + \beta||x^{\delta,h}(\alpha, \beta)||^2.$$

Lemma 4.1. *[6] For $\alpha, \beta \in \sum$, the partial derivatives of $F(\alpha, \beta)$ with respect to α and β are given by*

$$F'_\alpha(\alpha, \beta) = ||Bx^{\delta,h}(\alpha, \beta)||^2, \qquad F'_\beta(\alpha, \beta) = ||x^{\delta,h}(\alpha, \beta)||^2.$$

Proof. It is well known from the calculus that if some $u(\alpha) \in X$ is a differentiable function with respect to α and $u'(\alpha) \in X$ then

$$\begin{aligned}
\frac{d}{d\alpha}||u(\alpha)||^2 &= \lim_{t\to 0} \frac{\langle u(\alpha + t) - u(\alpha), u(\alpha)\rangle + \langle u(\alpha + t) - u(\alpha), u(\alpha + t)\rangle}{t} \\
&= \lim_{t\to 0}\langle \frac{1}{t}(u(\alpha + t) - u(\alpha)), u(\alpha)\rangle + \lim_{t\to 0}\langle \frac{1}{t}(u(\alpha + t) - u(\alpha)), u(\alpha + t)\rangle \\
&= \langle u'(\alpha), u(\alpha)\rangle + \langle u'(\alpha), u(\alpha)\rangle \\
&= 2\langle u(\alpha), u'(\alpha)\rangle.
\end{aligned}$$

(by symmetry of the scalar product). So, for $x := x^{\delta,h}(\alpha, \beta)$ we have

$$F'_\alpha(\alpha, \beta) = 2\langle A_h x, -y^\delta, A_h x'_\alpha\rangle + 2\alpha\langle Bx, Bx'_\alpha\rangle + ||Bx||^2 + 2\beta\langle x, x'_\alpha\rangle.$$

Using (4.1) with $g = x'_\alpha$ we have

$$\langle A_h x, -y^\delta, A_h x'_\alpha\rangle + \alpha\langle Bx, Bx'_\alpha\rangle + \beta\langle x, x'_\alpha\rangle = 0.$$

Thus, we end up with $F'_\alpha(\alpha, \beta) = ||Bx||^2$. □

In a similar way we can prove the lemma for F'_β.

Now with the help of Lemma 4.1, for $x^{\delta,h}(\alpha,\beta)$ we have

$$||A_h x - y^\delta||^2 = F(\alpha,\beta) - \alpha F'_\alpha(\alpha,\beta) - \beta F'_\beta(\alpha,\beta).$$

Therefore, the first equation in (3.11) can be rewritten as

$$F(\alpha,\beta) - \alpha F'_\alpha(\alpha,\beta) - \beta F'_\beta(\alpha,\beta) = (\delta + h\sqrt{F'_\beta(\alpha,\beta)})^2. \tag{4.2}$$

Once we approximate $F(\alpha,\beta)$ by a simple model function $m(\alpha,\beta)$, one can solve the corresponding approximate equation

$$m(\alpha,\beta) - \alpha m'_\alpha(\alpha,\beta) - \beta m'_\beta(\alpha,\beta) = (\delta + h\sqrt{m'_\beta(\alpha,\beta)})^2, \tag{4.3}$$

for α or β.

Remark 4.1. *For $g = x = x^{\delta,h}(\alpha,\beta)$ the variational form (4.1) gives*

$$||A_h x||^2 + \alpha||Bx||^2 + \beta||x||^2 = \langle A_h x, y^\delta\rangle.$$

So, for $x = x^{\delta,h}(\alpha,\beta)$,

$$\begin{aligned} F(\alpha,\beta) &= \langle A_h x - y^\delta, A_h x - y^\delta\rangle + \alpha||Bx||^2 + \beta||x||^2 \\ &= ||A_h x||^2 + ||y^\delta||^2 - 2\langle A_h x, y^\delta\rangle + \alpha||Bx||^2 + \beta||x||^2 \\ &= ||y^\delta||^2 - ||A_h x||^2 - \alpha||Bx||^2 - \beta||x||^2. \end{aligned}$$

The term $||A_h x||^2$ is approximated by $T||x||^2$, where T is a positive constant to be determined. Using this approximation together with Lemma 4.1 we end up with

$$F(\alpha,\beta) + \alpha F'_\alpha(\alpha,\beta) + (\beta + T)F'_\beta(\alpha,\beta) \approx ||y^\delta||^2.$$

A model function $m(\alpha,\beta)$ approximating $F(\alpha,\beta)$ can be found from differential equation

$$m(\alpha,\beta) + \alpha m'_\alpha(\alpha,\beta) + (\beta + T)m'_\beta(\alpha,\beta) = ||y^\delta||^2.$$

It can be checked that a simple parametric family of the solutions of this equation is given by

$$m(\alpha,\beta) = ||y^\delta||^2 + \frac{C}{\alpha} + \frac{D}{T+\beta}, \tag{4.4}$$

where C, D and T are constants to be determined. In the next section we will present an algorithm for the computation of the two regularization parameters α and β according to the dual RTLS rule

4.2 An algorithm for the approximate solution

So far we approximated the function $F(\alpha, \beta)$ by a model function $m(\alpha, \beta)$ to be able to solve (4.3). Now we present an algorithm for the approximate solution of the equations(3.11) by a special two-parameter model function approach.

Given $\alpha_0, \beta_0, y^\delta, A_h, \delta$ and h. Set $k := 0$.

1. Solve (3.10) with α_k, β_k to get $x^{\delta,h}(\alpha_k, \beta_k)$. Compute $F_1 = F(\alpha_k, \beta_k)$, $F_2 = F'_\alpha = ||Bx^{\delta,h}(\alpha_k, \beta_k)||^2$ and $F_3 = F'_\beta = ||x^{\delta,h}(\alpha_k, \beta_k)||^2$. In (4.4) set $C = C_k$, $D = D_k$, $T = T_k$ such that

$$\begin{cases} m(\alpha_k, \beta_k) = ||y^\delta||^2 + \frac{C}{\alpha_k} + \frac{D}{T+\beta_k} = F_1, \\ m'_\alpha(\alpha_k, \beta_k) = -\frac{C}{\alpha_k^2} = F_2, \\ m'_\beta(\alpha_k, \beta_k) = \frac{D}{(T+\beta_k)^2} = F_3. \end{cases}$$

   Then,

$$\begin{cases} C_k = -F_2\alpha_k^2, \\ D_k = -\frac{(||y^\delta||^2 - F_1 - F_2\alpha_k)^2}{F_3}, \\ T_k = \frac{||y^\delta||^2 - F_1 - F_2\alpha_k}{F_3} - \beta_k. \end{cases}$$

   Update $\beta = \beta_{k+1}$ using the second equation in (3.11) as

$$\beta_{k+1} = -\frac{h(\delta + h||x^{\delta,h}(\alpha_k, \beta_k)||)}{||x^{\delta,h}(\alpha_k, \beta_k)||},$$

   and update $\alpha = \alpha_{k+1}$ as the solution of the linear algebraic equation

$$m(\alpha, \beta_{k+1}) - \alpha m'_\alpha(\alpha, \beta_{k+1}) - \beta_{k+1} m'_\beta(\alpha, \beta_{k+1}) = (\delta + h\sqrt{m'_\beta(\alpha, \beta_{k+1})}).$$

   This equation is an approximate version of (3.11), (4.2), where $F(\alpha, \beta)$ is approximated by a model function $m(\alpha, \beta)$.

2. STOP if the stopping criteria $\max\left(\frac{|\alpha_{k+1} - \alpha_k|}{|\alpha_k|}, \frac{|\beta_{k+1} - \beta_k|}{|\beta_k|}\right) \leq \epsilon$ is satisfied; otherwise set $k := k + 1$, GO TO 1.

Though the convergence results for this algorithm are not known, it works well in experiments. Actually, It is not known under which conditions the algorithm is well defined. The only available result is: If the iteration converges, then the limit $(\alpha^*, \beta^*) = \lim_{k\to\infty}(\alpha_k, \beta_k)$ is a solution of the nonlinear system (3.11).

4.3 Dual Regularized Total Least Squares for Learning problem

In this section we discuss the dual RTLS for learning problem. In general we consider the linear operator equation

$$Ax = y, \tag{4.5}$$

together with the inequalities

$$||y - y^{\delta}|| \leq \delta,$$

and

$$||A - A_h|| \leq h.$$

Now we set $A = T$ and $A_h = T_x$.
That is,

$$||A - A_h|| = ||T - T_x||. \tag{4.6}$$

Then, in accordance to the discrete version $T_x f = S_x^* \mathbf{y}$ we set $y = \mathfrak{I}_{\mathcal{H}_K} f_0$ and $y^{\delta} = S_x^* \mathbf{y}$.

$$||y - y^{\delta}|| = ||\mathfrak{I}_{\mathcal{H}_k}^* f_0 - S_x^* \mathbf{y}||. \tag{4.7}$$

So, the dual RTLS problem (3.10) can be rewritten in the following form

$$(T_x^* T_x + \alpha B^* B + \beta) f = T_x^* S_x^* \mathbf{y}. \tag{4.8}$$

Now it is a time to define B^*B. From our numerical experience we set

$$B^* B = \sum_{i=1}^{n} K_{x_i} \langle K_{x_i}, . \rangle_{\mathcal{H}_K}. \tag{4.9}$$

Remark 4.2. *The operator T_x is self adjoint. That is, $T_x = T_x^*$.*

An important question which arises here is how do we choose h and δ? To answer this question we give the following lemma where the two quantity show up.

Lemma 4.2. *[1] Assume that the condition*

$$\int_Y (e^{\frac{|y - f_{\mathcal{H}}|}{M}} - \frac{|y - f_{\mathcal{H}}|}{M} - 1) \; dP(y|x) \leq \frac{\Sigma^2}{2M^2}$$

holds true, where $f_{\mathcal{H}}$ is the best approximation of f_0 from $\mathcal{H}$ with respect to $L^2(X, P(x)dx)$-norm, and Σ, $M \in \mathbb{R}$. Moreover, assume that

$$\sup_{x \in X} \sqrt{K(x,x)} \leq k < \infty$$

also holds true. For $0 < \eta \leq 1$ *and* $n \in \mathbb{N}$ *let*

$$G_\eta = \{z \in Z^n : ||T_x f_{\mathcal{H}} - S_x^* \boldsymbol{y}||_{\mathcal{H}} \leq \delta, ||T - T_x|| \leq h\}$$

with

$$\delta := \delta(n, \eta) = 2\{\frac{kM}{n} + \frac{k\Sigma}{\sqrt{n}}\} log \frac{4}{\eta},$$

$$h := h(n, \eta) = \frac{1}{\sqrt{n}} 2\sqrt{2} k^2 log \frac{4}{\eta}.$$

Then

$$P[G_\eta] \geq 1 - \eta.$$

To complete the answer of the above question, we give the following observation

Proposition 4.1. *From Lemma 4.2 and (4.6), (4.7) with probability* $1 - \eta$, $0 < \eta < 1$, *we have*

$$||y - y^\delta|| \leq \frac{c_1}{\sqrt{n}}, \quad ||T - T_x|| \leq \frac{c}{\sqrt{n}},$$

where c_1 *and* c *are two positive constants depending on* η, *and* n *is the size of the training set.*

Proof. We will show the first inequality; the second one can be obtained from Lemma 4.2 directly.
Since we have

$$T f_{\mathcal{H}} = \mathfrak{I}_{\mathcal{H}_k}^* f_0,$$

then,

$$\begin{aligned}
||y - y^\delta|| &= ||T f_{\mathcal{H}} - S_x^* \mathbf{y}|| \\
&= ||T f_{\mathcal{H}} - T_x f_{\mathcal{H}} + T_x f_{\mathcal{H}} - S_x^* \mathbf{y}|| \\
&\leq ||T f_{\mathcal{H}} - T_x f_{\mathcal{H}}|| + ||T_x f_{\mathcal{H}} - S_x^* \mathbf{y}|| \\
&\leq ||T - T_x|| ||f_{\mathcal{H}}|| + \delta \\
&\leq h||f_{\mathcal{H}}|| + \delta \\
&= \frac{1}{\sqrt{n}} 2\sqrt{2} k^2 log \frac{4}{\eta} ||f_{\mathcal{H}}|| + 2\{\frac{kM}{n} + \frac{k\Sigma}{\sqrt{n}}\} log \frac{4}{\eta} \\
&= \frac{1}{\sqrt{n}} [2\sqrt{2} k^2 log \frac{4}{\eta} ||f_{\mathcal{H}}|| + (2kMn^{\frac{-1}{2}} + 2kn\Sigma) log \frac{4}{\eta}] \\
&= \frac{c_1}{\sqrt{n}},
\end{aligned}$$

where $c_1 := 2\sqrt{2} k^2 log \frac{4}{\eta} ||f_{\mathcal{H}}|| + (2kMn^{\frac{-1}{2}} + 2kn\Sigma) log \frac{4}{\eta}$.

□

In our numerical experiments we took $c, c_1 \approx 0.1, 0.01$.

Proposition 4.2. *The neural network form*

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} c_i K(\boldsymbol{x}, \boldsymbol{x}_i), \tag{4.10}$$

is the minimizer of the problem (3.9) with B, A_h, y^δ given by (4.6), (4.7), (4.9) if and only if the coefficients c_i, $i = 1, \ldots, n$, satisfy the linear system

$$(n^2(\alpha \boldsymbol{K} + \beta \boldsymbol{I}) + \boldsymbol{K}^2)\boldsymbol{C} = \boldsymbol{K}\boldsymbol{y}. \tag{4.11}$$

where n is the size of the training set, α and β are the regularization parameters, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{C}= \{c_i\}_{i=1}^n$, $\boldsymbol{y}= \{\boldsymbol{y}_i\}_{i=1}^n$ and $\boldsymbol{K}$ is the kernel matrix, namely

$$\begin{pmatrix} K(x_1,x_1) & K(x_1,x_2) & \ldots & K(x_1,x_n) \\ K(x_2,x_1) & K(x_2,x_2) & \ldots & K(x_2,x_n) \\ \vdots & \vdots & & \vdots \\ K(x_n,x_1) & K(x_n,x_2) & \ldots & K(x_n,x_n) \end{pmatrix}.$$

Proof. Substituting from (2.10), (4.9) and (4.10) into (4.8) we get
$\frac{1}{n^2} \sum_{i=1}^{n} K_{x_i} \langle K_{x_i}, . \rangle_{\mathcal{H}_K} \sum_{j=1}^{n} K_{x_j} \langle K_{x_j}, \sum_{l=1}^{n} c_l K_{x_l} \rangle_{\mathcal{H}_K} \; + \; \alpha \sum_{j=1}^{n} K_{x_j} \langle K_{x_j}, \sum_{l=1}^{n} c_l K_{x_l} \rangle_{\mathcal{H}_K} \; +$
$\beta \sum_{j=1}^{n} c_j K_{x_j} \;=\; \frac{1}{n^2} \sum_{i=1}^{n} K_{x_i} \langle K_{x_i}, \sum_{j=1}^{n} y_j K_{x_j} \rangle_{\mathcal{H}_K}.$

Multiplying both sides by n^2 and equating the coefficients of K we get

$$\sum_{j=1}^{n} \sum_{i,l=1}^{n} c_l \langle K_{x_i}, K_{x_j} K(x_j, x_l) \rangle_{\mathcal{H}_K} \; + \; n^2 \alpha \sum_{l=1}^{n} c_l K(x_j, x_l) \; + \; n^2 \beta \sum_{j=1}^{n} c_j \;=\; \sum_{i=1}^{n} \sum_{j=1}^{n} y_j K(x_i, x_j)$$

$$\Longleftrightarrow$$

$$\begin{pmatrix} K(x_1,x_1) & \ldots & K(x_1,x_n) \\ \vdots & & \vdots \\ K(x_n,x_1) & \ldots & K(x_n,x_n) \end{pmatrix} \begin{pmatrix} K(x_1,x_1) & \ldots & K(x_1,x_n) \\ \vdots & & \vdots \\ K(x_n,x_1) & \ldots & K(x_n,x_n) \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} +$$
$$n^2\alpha \begin{pmatrix} K(x_1,x_1) & \ldots & K(x_1,x_n) \\ \vdots & & \vdots \\ K(x_n,x_1) & \ldots & K(x_n,x_n) \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} + n^2\beta \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} =$$
$$\begin{pmatrix} K(x_1,x_1) & \ldots & K(x_1,x_n) \\ \vdots & & \vdots \\ K(x_n,x_1) & \ldots & K(x_n,x_n) \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Thus,

$$\mathbf{K}^2\mathbf{C} + n^2\alpha\mathbf{K}\mathbf{C} + n^2\beta\mathbf{C} = \mathbf{K}\mathbf{y}.$$

or as written in (4.11)

$$(n^2(\alpha\mathbf{K} + \beta\mathbf{I}) + \mathbf{K}^2)\mathbf{C} = \mathbf{K}\mathbf{y}.$$

□

The conclusion that can be drawn from the present section is that using a regularization network of the form (4.10), for a certain class of kernels K, is equivalent to minimizing functionals of the form (3.9). It should be mentioned that the choice of K is equivalent to the choice of a corresponding RKHS. In Figure 4.1 a variety of kernels widely used is listed.

Despite the importance of determining the coefficients c_i , $i = 1, \ldots, n$, for the neural network form of the minimizer, the choice of the two regularization parameters α and β is more important and, actually, crucial.

Kernel function	Regularization Network
$K(\mathbf{x}-\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2)$	Gaussian RBF
$K(\mathbf{x}-\mathbf{y}) = (\|\mathbf{x}-\mathbf{y}\|^2 + c^2)^{-1/2}$	Inverse multiquadric
$K(\mathbf{x}-\mathbf{y}) = (\|\mathbf{x}-\mathbf{y}\|^2 + c^2)^{1/2}$	Multiquadric
$K(\mathbf{x}-\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^{2n+1}$	Thin plate splines
$K(\mathbf{x}-\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^{2n} \ln(\|\mathbf{x}-\mathbf{y}\|)$	
$K(\mathbf{x},\mathbf{y}) = \tanh(\mathbf{x}\cdot\mathbf{y}-\theta)$	(only for some values of θ) Multi-Layer Perceptron
$K(\mathbf{x},\mathbf{y}) = (1+\mathbf{x}\cdot\mathbf{y})^d$	Polynomial of degree d
$K(x,y) = B_{2n+1}(x-y)$	B-splines
$K(x,y) = \dfrac{\sin(d+1/2)(x-y)}{\sin((x-y)/2)}$	Trigonometric polynomial of degree d

Figure 4.1: Some possible kernel functions

4.4 Prediction scheme

In Figure 4.2 we present the algorithm discussed above.

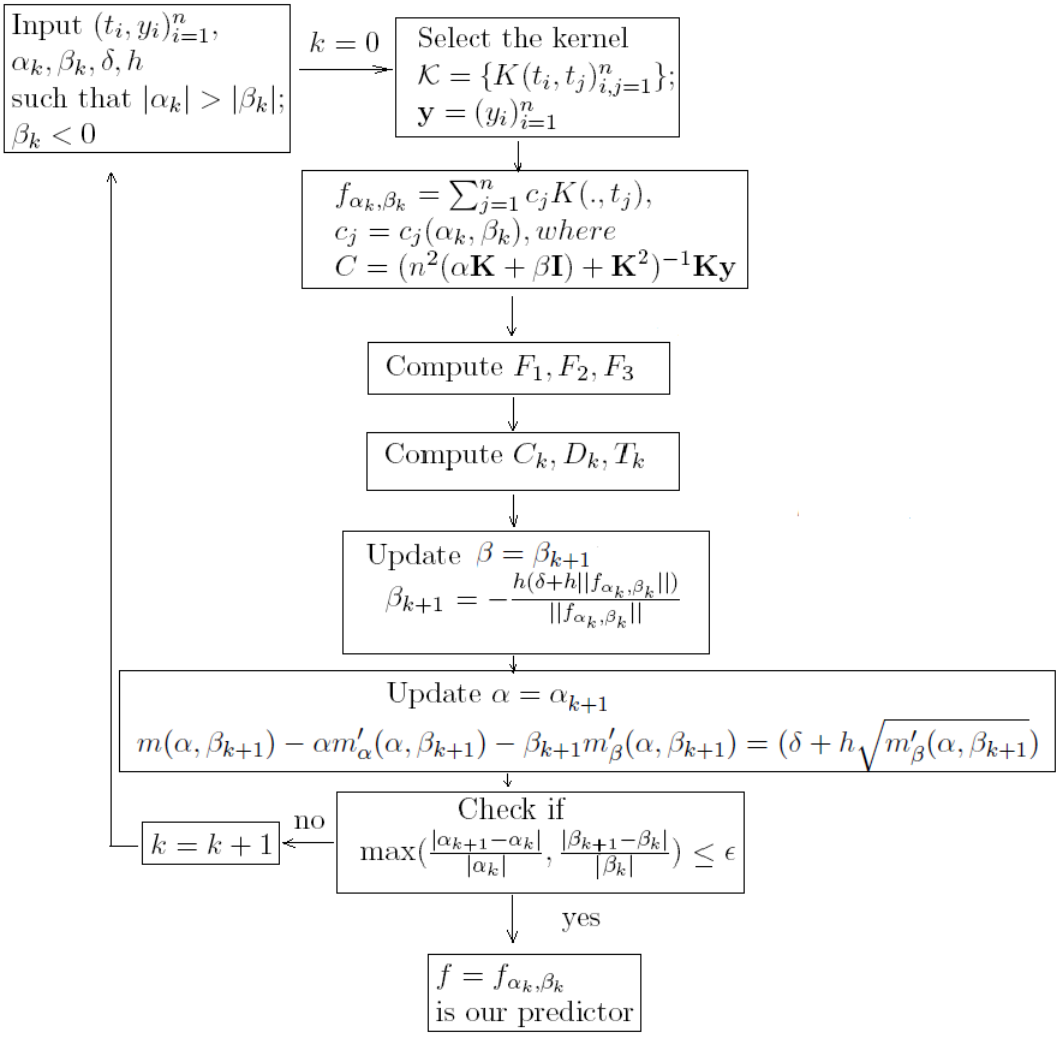

Figure 4.2: An algorithm process to choose α and β to get our predictor.

Chapter 5

Numerical Examples

The purpose of this chapter is to demonstrate several results from the implementations of the theoretical framework applied to some problems. First of all, we should distinguish here two different types of prediction. The first type happens when inputs from the training set can be seen as boundary points of some area and all further inputs are expected to appear in this area, this is the so-called "interpolation " type. However, in many application a new input will appear outside of the area, then we will talk about "extrapolation". Actually, interpolation is an important feature of Learning Theory; it is the procedure to estimate values at unknown locations within the area covered by existing observations. On the other hand, many areas of mathematics, statistics, and computer science deal with extrapolation of functions from partial information or examples.

In the current chapter we will discuss both the interpolation and extrapolation types within two test examples which have been introduced in [3]. In both examples we try to learn a target function $f : [0, 2\pi] \to \mathbb{R}$ from a set of its samples. As mentioned, the target functions are always unknown, but a discrete training data can be measured in a certain frequency in a time periode τ. When given a training data, our algorithm is expected to give a reasonable estimator/ predictor.

• **Example 1.** The best predictor, or the target function is $f(t) = \frac{1}{10}(t + 2(e^{-8(\frac{4\pi}{3}-t)^2} - e^{-8(\frac{\pi}{2}-t)^2} - e^{-8(\frac{3\pi}{2}-t)^2}))$, $t \in [0, 2\pi]$. This function belongs to RKHS generated by the kernel $K(x,t) = xt + e^{-8(t-x)^2}$ which will be used in the framework of learning algorithm based on dual RTLS . First of all, we generate the training set by sampling the target function. The following training set τ_n^m consists of $m + 1$ points and is obtained by taking sample frequency as $2\pi/n$, and adding to each discrete value the random noise in a range of $[-0.02, 0.02]$.

i	0	1	2	3	4	5	6
t_i	0	0.3142	0.6283	0.9425	1.2566	1.5708	1.8850
$f(t_i)$	0.0126	0.0476	0.0477	0.1023	0.0401	-0.0590	0.0888

i	7	8	9	10	11	12	13
t_i	2.1991	2.5133	2.8274	3.1416	3.4558	3.7699	4.0841
$f(t_i)$	0.2133	0.2695	0.3013	0.3005	0.3671	0.4443	0.5825

i	14	15	16	17	18	19	20
t_i	4.3982	4.7124	5.0265	5.3407	5.6549	5.9690	6.2832
$f(t_i)$	0.5018	0.2792	0.4094	0.5422	0.5770	0.6153	0.6345

Table 5.1: Training set τ_{20}^{20} for f from Example 1.

5.1 Interplation type prediction for Example 1

In this section we discuss the interpolation type prediction for the target function introduced in Example 1 using both Tikhonov technique and our new approach based on the dual RTLS.

One can successfully predict the value of the tested function $f(t)$ at any point of the interval $[0, 2\pi]$ containing all inputs from the training set τ_{20}^{20} using Tikhonov regularization technique. This can be seen clearly in Figure 5.1 [taken from [7]].

Meanwhile, using the dual RTLS approch, one can also successfully predict the value of the same tested function at any point of the same interval containing all inputs from the training set based, again, on τ_{20}^{20}, as it is shown in Figure 5.2.

So, we conclude that both regularization techniques, Tikhonov and the dual RTLS, are very good in the context of interpolation type of the prediction.

5.2 Extrapolation type prediction for Example 1

As we mentioned, Tikhonov regularization technique does not give satisfactory results when predicting the value of the tested function in some situation. What we exactly meant by “some situations” is the extrapolation type of the prediction.

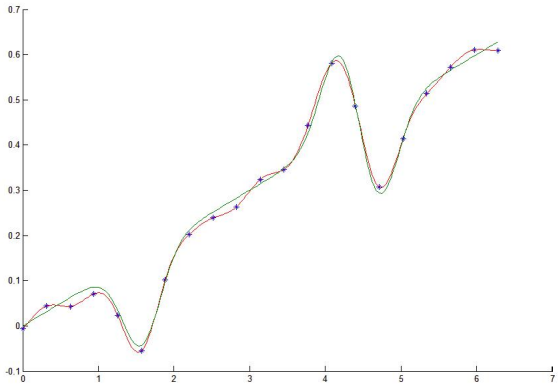

Figure 5.1: Prediction for inputs within the scope of training set of 21 points: ideal predictor (green line) and its approximation given by Tikhonov learning algorithm based on the kernel $K(x,t) = xt + e^{-8(t-x)^2}$ (red line).

That is, we show that using a one parameter regularization method, namely Tikhonov, does not give a good prediction using a part of the data. This was shown in a recent research [7]. Figure 5.3 [taken from [7]] shows that the quality of the prediction at the points of the interval $[0, 2\pi]$ being beyond the scope of the training set $\tau_{20}^{15} = \{t_i = \pi i/10, i = 0, 1, \ldots, 15\}$ is rather poor.

On the other hand, when a prediction beyond the scope of the training set based on τ_{20}^{15} using the dual RTLS technique, we can successfully predict the value of the tested function as it is seen in Figure 5.4.
In Figure 5.5 we present the prediction using the dual RTLS made for the training set based on τ_{50}^{45} which looks promising.

• **Example 2.** The best predictor, or the target function is $f(t) = sin(t) + \frac{1}{2}sint(3t), \quad t \in [0, 2\pi]$, This function belongs to RKHS generated by the kernel $K(x,t) = \frac{2}{3}sin(x)sin(t) + \frac{1}{3}sin(3x)sin(3t), t \in [0, 2\pi]$. As we mentioned before, we generate the training set by sampling the target function by taking sample frequency $2\pi/20$ and adding to each discrete value the random noise in range of $[-0.2, 0.2]$, which is different from the range used in Example 1. we obtained the training set shown in Table 5.2.

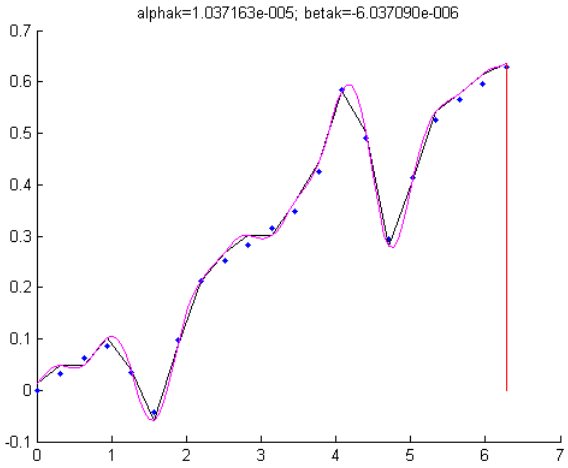

Figure 5.2: (Example 1) Prediction within the scope of training set of 21 points: Blue dots present the ideal predictor, magenta line presents its approximation given by dual RTLS learning algorithm and the black line presents the noisy data.

5.3 Interpolation type prediction for Example 2

Exactly as in Example 1, one can successfully predict the value of the tested function $f(t)$ using the dual RTLS technique at any point in the interval $[0, 2\pi]$ which contains all input within the scope of τ_{20}^{20}. The quality of the prediction is shown in Figure 5.6.

One can also predict successfully the value of the tested function $f(t)$ using Tikhonov regularization at any point in the interval $[0, 2\pi]$ which contains all input within the scope of τ_{20}^{20} [see Experiment 2 in [3]].

5.4 Extrapolation type prediction for Example 2

It can be seen clearly in Figure 5.7 that the quality of the prediction of the tested function is very good using the same training data $\tau_{15}^{20}$and the dual RTLS for the construction of the prediction. The prediction here is made beyond the scope of the training set.

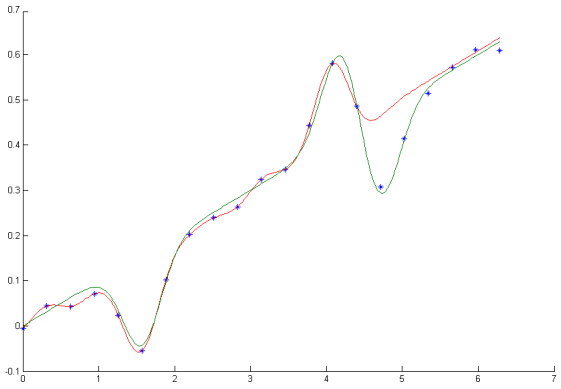

Figure 5.3: Prediction for inputs beyond the scope of training set of 16 points: ideal predictor (green line) and its approximation given by Tikhonov learning algorithm based on the kernel $K(x,t) = xt + e^{-8(t-x)^2}$ (red line).

In Figure 5.8 one can see the prediction given by the dual RTLS beyond the scope of the set τ_{50}^{45}.

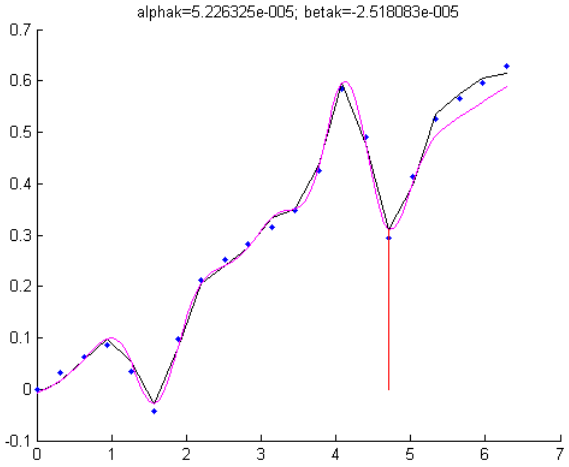

Figure 5.4: (Example 1) Prediction beyond the scope of training set of 16 points: Blue dots present the ideal predictor, magenta line presents its approximation given by dual RTLS learning algorithm and the black line presents the noisy data.

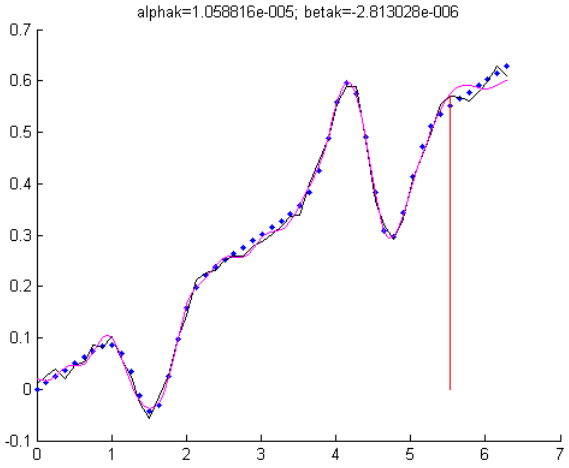

Figure 5.5: (Example 1) Prediction beyond the scope of training set of 46 points: Blue dots present the ideal predictor, magenta line presents its approximation given by dual RTLS learning algorithm and the black line presents the noisy data.

i	0	1	2	3	4	5	6
t_i	0	0.3142	0.6283	0.9425	1.2566	1.5708	1.8850
$f(t_i)$	-0.1857	0.8532	1.2369	1.0350	0.7603	0.5973	0.6141

i	7	8	9	10	11	12	13
t_i	2.1991	2.5133	2.8274	3.1416	3.4558	3.7699	4.0841
$f(t_i)$	1.0257	0.9318	0.7959	-0.1873	-0.8028	-1.2448	-1.1247

i	14	15	16	17	18	19	20
t_i	4.3982	4.7124	5.0265	5.3407	5.6549	5.9690	6.2832
$f(t_i)$	-0.5278	-0.4221	-0.7303	-0.7834	-1.2495	-0.7380	-0.0474

Table 5.2: Training set τ_{20}^{20} for f from Example 2.

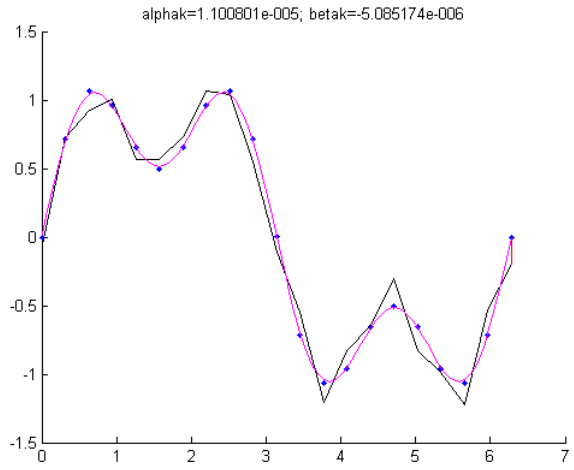

Figure 5.6: (Example 2) Prediction within the scope of training set of 21 points: Blue dots present the ideal predictor, magenta line presents its approximation given by dual RTLS learning algorithm and the black line presents the noisy data.

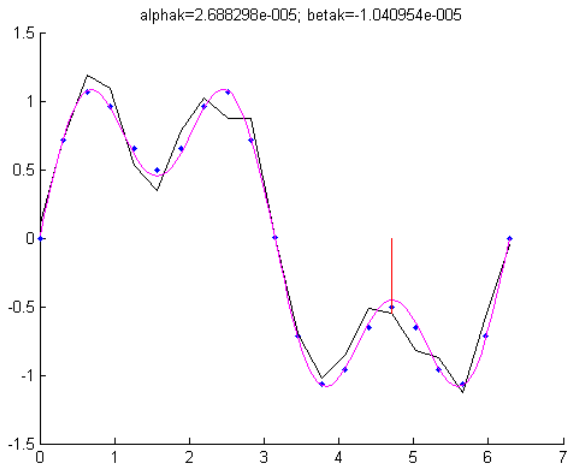

Figure 5.7: (Example 2) Prediction beyond the scope of training set of 16 points: Blue dots present the ideal predictor, magenta line presents its approximation given by dual RTLS learning algorithm and the black line presents the noisy data.

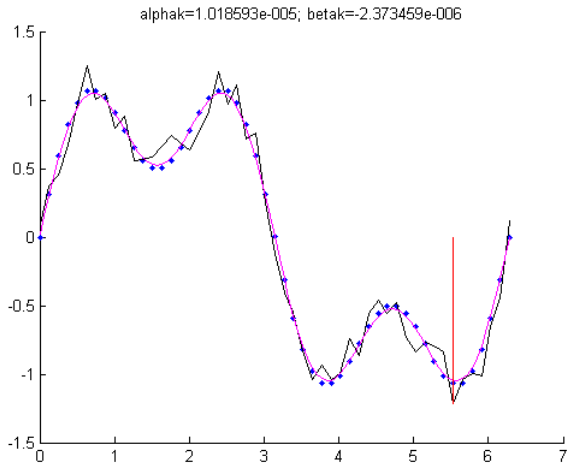

Figure 5.8: (Example 2) Prediction beyond the scope of training set of 46 points: Blue dots present the ideal predictor, magenta line presents its approximation given by dual RTLS learning algorithm and the black line presents the noisy data.

Chapter 6

Conclusions

There is a strong connection between Learning Theory and inverse problems. The problem of learning can be reduced to an ill-posed linear operator equation in the reproducing kernel Hilbert space (RKHS). For solving this ill-posed operator equation, we considered the calssical Tikhonov regularization and the newly introduced dual regularized toltal least squares (dual RTLS). It was known that Tikhonov regularization applied to the equations coming from prediction problems is not satisfactory. We showed that the application of the dual RTLS to such equations leads to much better results.

In the dual RTLS it is crucial to choose the two regularization parameters which show up in the method. For this reason we have constructed a selection algorithm which is able to determine the optimal parameters.

The performance of the constructed algorithm was tested on test examples taken from the literature. From the presented numerical results it can be seen that

1. The interpolation type prediction of the tested function can be done in a very good way using both Tikhonov regularization and the dual RTLS.

2. The extrapolation type prediction of the tested function is rather poor if Tikhonov regularization is employed. However, using the dual RTLS to perform this type of prediction is very good and looks promising.

Bibliography

[1] F. Bauer, S. Pereverzyev, L. Rosasco, On regularization algorithms in learning theory, Journal of Complexity 23(2007) 52-72.

[2] T. Evgeniou, T. Poggio, M. Pontil, Regularization Network and Support Vector Machines, Advances in Computational Mathematics 13(2000)1-50.

[3] C.A. Micchelli, M. Pontil, Learning the Kernel Function via Regularization, Journal of Machine Learning Research 6(2005) 1099-1125.

[4] T. Poggio, S. Smale, The Mathematics of Learning: Dealing with Data, Notice of the AMS, Volume 50, Number5, 2003.

[5] S. Lu, S. Pereverzyev, U.Tautenhahn, Regularized toltal least squares:Compuational aspects and error bounds, ricam report number 2007-30.

[6] S. Lu, S. Pereverzyev, U.Tautenhahn, A model function method in toltal least squares, ricam report number 2008-18.

[7] Hujun Wang, Adaptive regularization in Learning Theory; Case study : Prediction Of Blood Glucose Level, Master dissertation, Linz, Austria, 2008.

[8] S.Lu, S.V.Pereverzev, U.Tautenhahn, Dual regularized total least squares and multi-parameter regularization, Computational Methods in Applied Mathematics, Vol.8(2008), p. 253-262.

[9] V.N. Vapnik, Statistical Learning Theory (Wiley, New York, 1998).

[10] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 686 (1950) 337 -404.

[11] Golub, G. H., Hansev, P. C. and O'Leary, D. P. (1999): Tikhonov regularization and total least squares, SIAM J. Matrix Anal. 21, 185 -194.

[12] Huffel, S. V. and Vanderwalle, J. (1991): The Total Least Squares Problem: Computional Aspects and Analysis, Philadelphia: SIAM.

[13] Renaut, R. A. and Guo, H. (2005): Efficient algorithms for solutionsof regularized total least squares, SIAM J. Matrix Anal. Appl. 26, 457 - 476.

[14] Sima, D., Huffel, S. V. and Golub, G. H. (2004): Regularized total least squares based on quadratic eigenvalue problem solvers, BIT Numerical Mathematics 44, 793 - 812.

[15] Beck, A., Ben-Tal, A. and Teboulle, M. (2006): Finding a global optimal solution for a quadratically constrained fractional problem with applications to the regularized total least squares, SIAM J. Matrix Anal. Appl. 28, 425 - 445.

[16] Ivanov, V. K. (1966): On the approximate solution of operator equations of the first kind, USSR Comp. Math. Math. Phys. 6, 1089 - 1094.

[17] Kunisch, K. and Zou, J. (1998): Iterative choice of regularization parameters in linear inverse problems, Inverse Problems 14, 1247-1264.