

Chapter 15

Multiparameter Regularization for Construction of Extrapolating Estimators in Statistical Learning Theory

Shuai Lu, Sergiy Pereverzyev Jr., and Sivananthan Sampath

Abstract One-parameter regularization methods, such as the Tikhonov regularization, are used to solve the operator equation for the estimator in the statistical learning theory. Recently, there has been a lot of interest in the construction of the so called extrapolating estimators, which approximate the input–output relationship beyond the scope of the empirical data. The standard Tikhonov regularization produces rather poor extrapolating estimators. In this paper, we propose a novel view on the operator equation for the estimator where this equation is seen as a perturbed version of the operator equation for the ideal estimator. This view suggests the dual regularized total least squares (DRTLS) and multi-penalty regularization (MPR), which are multi-parameter regularization methods, as methods of choice for constructing better extrapolating estimators. We propose and test several realizations of DRTLS and MPR for constructing extrapolating estimators. It will be seen that, among the considered realizations, a realization of MPR gives best extrapolating estimators. For this realization, we propose a rule for the choice of the used regularization parameters that allows an automatic selection of the suitable extrapolating estimator.

S. Lu (✉)

School of Mathematical Science, Fudan University, 200433 Shanghai, China
e-mail: slu@fudan.edu.cn

S. Pereverzyev Jr.

Industrial Mathematics Institute, Johannes Kepler University of
Linz, Altenberger Strasse 69, 4040 Linz, Austria
e-mail: pereverzyev@indmath.uni-linz.ac.at

S. Sampath

Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy
of Sciences, Altenberger Strasse 69, 4040 Linz, Austria
e-mail: sivananthan.sampath@oeaw.ac.at

15.1 Introduction

Let us consider a system as a functioning entity that takes an input and gives the output. In many scientific studies, one would like to understand how a specific system performs, i.e., given an input how the system produces the output. In particular, one would like to be able to predict the system output. It is very difficult to access the internal structure of many systems, and this complicates the discovery of the system internal functioning mechanisms. In this case, the available information about the system are input–output pairs, which are obtained from the system observations and often called the *empirical data*.

In machine learning [1, 7, 27], a part of computer science, one is concerned with the design and development of algorithms, called (machine) learning algorithms, that allow computers (machines) to predict (to make a decision about) the system output based on the empirical data from the system observations.

The analysis of learning algorithms is done in the framework of (computational) learning theories. One of such theories is the so-called statistical learning theory [30, 33]. According to this theory, the learning algorithm should construct a function, called an *estimator*, that approximates well the relationship between system input and system output. The theory defines the measure of the approximation quality of an estimator and, according to this measure, an *ideal estimator* that has the best approximation quality over a specified function space.

Usually, the ideal estimator cannot be constructed. So, the task of the learning algorithm is to use the empirical data for constructing an estimator that converges to the ideal estimator when the number of observations goes to infinity. The theory suggests a natural approach for constructing an estimator based on the empirical data. This approach leads to an operator equation for the estimator.

As it was observed in [14, 20], there is a similarity between the construction of an estimator and the solution of inverse problems, which are usually formulated as operator equations [15, 17, 18, 31]. Many inverse problems are ill-posed, and for their stable solution, one uses the so-called regularization methods. The operator equation for the estimator in the statistical learning theory is also ill-posed: it does not have a unique solution, and many solutions of this equation are far away from the desired ideal estimator. So, this suggests to apply the regularization methods from the theory of inverse problems. In [16, 29], it was proposed to use the Tikhonov regularization method for solving the operator equation for the estimator. Application of general regularization methods, which are used for solving ill-posed inverse problems, is considered in [4].

One can distinguish between two types of estimators: interpolating and extrapolating. In the case of the interpolating estimator, the inputs in the empirical data are coming from some specified set, and further inputs are also expected to come from this set. One can also say that the interpolating estimator provides a prediction at the unknown inputs within the set that is defined by the existing observations. Whereas, the extrapolating estimator provides a prediction outside this set.

It has been observed that Tikhonov regularization could give good interpolating estimators. On the contrary, the extrapolating estimators that are constructed by the Tikhonov regularization have a rather poor approximation quality. Thus, alternative methods for constructing extrapolating estimators are needed.

Our analysis of the operator equation for the estimator suggests that it can be viewed as a perturbed version of the operator equation for the ideal estimator where both the operator and the right-hand side are modified (perturbed). Recently, in the regularization theory, there has been developed a method, called the dual regularized total least squares (DRTLS) [23–25], which is designed for perturbed operator equations. Therefore, this method can be suggested to solve the operator equation for the estimator. For each realization of DRTLS one can construct a corresponding realization of the so-called multi-penalty regularization (MPR) [8, 21] method. This method can be also suggested to solve the operator equation for the estimator.

Tikhonov regularization belongs to a family of the so-called one-parameter regularization methods. On the contrary, DRTLS and MPR are multiparameter regularization methods. This gives them a bigger flexibility for the solution of the perturbed operator equations. And so, one could expect that they could construct better extrapolating estimators.

In this chapter, for solving the operator equation for the estimator, we propose several realizations of DRTLS and MPR. The quality of the extrapolating estimators that are constructed by these realizations will be compared. It will turn out that, among the considered realizations, a realization of MPR gives best extrapolating estimators.

Each realization of a regularization method requires a rule for the choice of the regularization parameters that are used in the method. We will propose such a rule for the mentioned realization of MPR that constructs best extrapolating estimators.

This chapter is organized as follows. In Sect. 15.2, we review the main concepts of the statistical learning theory and derive the operator equation for the estimator. DRTLS and MPR are presented in Sect. 15.3. The perturbation levels in the operator equation for the estimator, which can be used in the application of regularization methods, are estimated in Sect. 15.4. We present several realizations of DRTLS and MPR as well as the comparison of extrapolating estimators that are obtained by these realizations in Sect. 15.5. For the realization that gives the best extrapolating estimators, we propose a rule for the choice of the used regularization parameters in Sect. 15.6. This chapter is finished with conclusions and outlook in Sect. 15.7.

15.2 The Problem of the Construction of an Estimator in the Statistical Learning Theory

In the statistical learning theory, the empirical data $\mathbf{z} = \{(x_i, y_i), i = 1, \dots, n\}$ are seen as the realizations of random variables $(x, y) \in X \times Y$ with a probability density $p(x, y)$. Specifically, we consider the situation when both X and Y are subsets of \mathbb{R} .

One of the central problems in the statistical learning theory is the construction of the estimator $f : X \rightarrow Y$ that approximates well the relationship between x and y , i.e. $y \approx f(x)$. The common way of measuring the approximation quality of f is the consideration of the expected error:

$$\mathcal{E}^2(f) = \int_{X \times Y} (y - f(x))^2 p(x, y) dx dy.$$

Minimization of $\mathcal{E}(f)$ over an appropriate function space leads to an *ideal estimator*. A rather broad function space, for which it is also possible to give an explicit form of the corresponding ideal estimator, is obtained from the following splitting of the density p :

$$p(x, y) = p_x(x) p_{y|x}(y|x), \tag{15.1}$$

where $p_x(x) = \int_Y p(x, y) dy$ is the so-called marginal probability density, and $p_{y|x}(y|x)$ is the so-called conditional probability density for y given x . Then, the minimizer of $\mathcal{E}(f)$ over the function space

$$L^2(X, p_x) = \left\{ f : X \rightarrow Y \mid \|f\|_p^2 := \int_X f^2(x) p_x(x) dx < +\infty \right\}$$

is given by

$$f_p(x) = \int_Y y p_{y|x}(y|x) dy.$$

Minimization of the expected error $\mathcal{E}(f)$ over a subspace $\mathcal{H} \subset L^2(X, p_x)$, i.e.,

$$\mathcal{E}(f) \rightarrow \min_{f \in \mathcal{H}}, \tag{15.2}$$

means in fact finding a function $f \in \mathcal{H}$ that best approximates $f_p(x)$ in $L^2(X, p_x)$, i.e., a function for which the norm $\|f - f_p\|_p^2$ is minimal. This follows from the following property of the expected error:

$$\mathcal{E}^2(f) = \|f - f_p\|_p^2 + \mathcal{E}^2(f_p).$$

This fact allows the formulation of the operator equation for the solution of (15.2). Let $\mathcal{J} : \mathcal{H} \rightarrow L^2(X, p_x)$ be the inclusion operator. Then, solution of (15.2) satisfies the operator equation

$$\mathcal{J}^* \mathcal{J} f = \mathcal{J}^* f_p. \tag{15.3}$$

It is common to assume that this equation is uniquely solvable and define its solution as f^\dagger [13, 14].

Since the probability density p is usually unknown in practice, function f^\dagger provides an ideal estimator that one cannot have but that one tries to approximate using the empirical data \mathbf{z} . In the construction of this approximating estimator $f_{\mathbf{z}}$, an important role is played by the so-called *empirical error*

$$\mathcal{E}_{\text{emp}}^2(f) = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2,$$

which is the statistical approximation of the expected error. The first idea for the construction of the estimator $f_{\mathbf{z}} \in \mathcal{H}$ could be to find such $f_{\mathbf{z}}$ that minimizes the empirical error over the function space \mathcal{H} , i.e., to solve the following minimization problem:

$$\mathcal{E}_{\text{emp}}(f) \rightarrow \min_{f \in \mathcal{H}}. \quad (15.4)$$

However, usually there are many minimizers of $\mathcal{E}_{\text{emp}}(f)$, even such that $\mathcal{E}_{\text{emp}}(f) = 0$, but among them, there are many that are far away from the desired f^\dagger .

Before discussing the further steps, let us formulate an operator equation for the minimizer of (15.4). For this purpose, let us define the so-called *sampling operator* $S_{\mathbf{x}} : \mathcal{H} \rightarrow \mathbb{R}^n$ that acts as follows $S_{\mathbf{x}} : f \mapsto (f(x_1), f(x_2), \dots, f(x_n))$ and take the following weighted euclidean norm in \mathbb{R}^n : $\|\mathbf{x}\|^2 = \frac{1}{n} \sum_{i=1}^n x_i^2$ for $\mathbf{x} \in \mathbb{R}^n$. Then, the empirical error can be written as

$$\mathcal{E}_{\text{emp}}(f) = \|\mathbf{y} - S_{\mathbf{x}}f\|,$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)$. And so, the minimization problem (15.4) is equivalent to solving the operator equation

$$S_{\mathbf{x}}^* S_{\mathbf{x}} f = S_{\mathbf{x}}^* \mathbf{y}. \quad (15.5)$$

As the minimization problem (15.4), also the operator equation (15.5) does not have a unique solution, and there are many solutions of (15.5) that are far away from f^\dagger .

In [16, 29], it was proposed to modify (15.4) using the Tikhonov regularization:

$$\mathcal{E}_{\text{emp}}^2(f) + \beta \|f\|_{\mathcal{H}}^2 \rightarrow \min_{f \in \mathcal{H}}, \quad (15.6)$$

where $\beta > 0$ is the so-called regularization parameter. The minimization problem (15.6) is equivalent to solving the following operator equation:

$$(S_{\mathbf{x}}^* S_{\mathbf{x}} + \beta I)f = S_{\mathbf{x}}^* \mathbf{y}, \quad (15.7)$$

where $I : \mathcal{H} \rightarrow \mathcal{H}$ is the identity operator. The regularization parameter β has to be chosen such that the corresponding estimator, i.e., the solution of (15.6) or (15.7), approximates well the ideal estimator f^\dagger .

As it was mentioned in the Introduction, two situations can be distinguished. In the first situation, the further inputs x are expected to come from a set X_{ϵ} that is

defined by the existing inputs $\mathbf{x} = (x_1, x_2, \dots, x_n)$. This set X_e is usually $\text{conv}\{x_i, i = 1, \dots, n\}$. The estimators that correspond to this situation are called interpolating estimators. It is quite well-known that the regularization parameter β in (15.7) can be chosen such that the solution of (15.7) is a good interpolating estimator, i.e., it approximates well the ideal estimator f^\dagger .

On the contrary, in the second situation, the further inputs x are expected to come also outside X_e . The estimators in this situation are called extrapolating estimators. As it will be seen in Sect. 15.5, estimators that are constructed by Tikhonov regularization, i.e., the solutions of (15.7) for various values of β , have bad extrapolating properties. Thus, for the construction of the extrapolating estimators, other methods are needed.

To our best knowledge, it has not been yet observed that equation (15.5) can be viewed as a perturbed version of the operator equation (15.3). As it will be seen in Sect. 15.4, as the number of observations n increases, the operator and the right-hand side in (15.5) approach the operator and the right-hand side in (15.3). More precisely, in corresponding norms, it holds that

$$\begin{aligned} \lim_{n \rightarrow \infty} \| \mathcal{J}^* \mathcal{J} - S_{\mathbf{x}}^* S_{\mathbf{x}} \| &= 0, \\ \lim_{n \rightarrow \infty} \| \mathcal{J}^* f_p - S_{\mathbf{x}}^* \mathbf{y} \| &= 0. \end{aligned}$$

Such a view suggests that the operator equation (15.5) can be treated by the recently developed DRTLS method [23–25] and the corresponding MPR [8, 21] method.

15.3 Dual Regularized Total Least Squares and Multi-penalty Regularization

Let us assume that there is an operator $A_0 : \mathcal{F} \rightarrow \mathcal{G}$, which acts between Hilbert spaces \mathcal{F} , space of solutions, and \mathcal{G} , space of data. Assume further, that for some perfect data $g_0 \in R(A_0) \subset \mathcal{G}$, there is a unique solution $f_0 \in \mathcal{F}$ to the problem

$$A_0 f = g_0. \tag{15.8}$$

Now, consider the situation when the pair (A_0, g_0) is not known, but instead, we are given an operator $A_h : \mathcal{F} \rightarrow \mathcal{G}$ and data $g_\delta \in \mathcal{G}$ that can be seen as noisy versions of the operator A_0 and data g_0 such that

$$\begin{aligned} \|g_0 - g_\delta\| &\leq \delta, \\ \|A_0 - A_h\| &\leq h, \end{aligned}$$

with some known noise levels $\{\delta, h\} \subset (0, +\infty)$.

For the ill-posed problem (15.8), the operator equation

$$A_h f = g_\delta \tag{15.9}$$

may have no solution, or its solution may be arbitrarily far away from f_0 . In this case, the so-called regularization methods [15, 17, 18, 31] are used. Many regularization methods consider the situation when only the data has some noise, and the involved operator A_0 is known exactly. A method, called DRTLS, that takes into account also the noise in the operator has been recently proposed in [23–25]. We review this method below.

Let us fix an operator B that is defined on \mathcal{F} and acts to some other Hilbert space. The idea of DRTLS is to approximate f_0 by the solution of the following minimization problem:

$$\|Bf\| \rightarrow \min_{f,A} \quad \text{subject to} \quad \|A - A_h\| \leq h, \|Af - g_\delta\| \leq \delta. \tag{15.10}$$

The solution of this minimization problem for which its constrains are active solves the following operator equation:

$$(A_h^* A_h + \alpha B^* B + \beta I) f = A_h^* g_\delta, \tag{15.11}$$

where $I : \mathcal{F} \rightarrow \mathcal{F}$ is the identity operator and α, β satisfy the following conditions:

$$\begin{aligned} \|A_h f_{\alpha,\beta} - g_\delta\| &= \delta + h \|f_{\alpha,\beta}\|, \\ \beta &= -\frac{h(\delta + h \|f_{\alpha,\beta}\|)}{\|f_{\alpha,\beta}\|}, \end{aligned} \tag{15.12}$$

where $f_{\alpha,\beta}$ is the solution of the operator equation (15.11) for the fixed α, β .

An iterative procedure for approximating the pair (α, β) in (15.12) has been proposed in [25]. It should be noted that $\beta < 0$ in (15.12). On the other hand, the operator equation (15.11) with $\alpha > 0$ and $\beta > 0$ arises in the application of the so-called MPR (see, e.g., [8, 21]) to the operator equation (15.9), where the following minimization problem is considered:

$$\|A_h f - g_\delta\|^2 + \alpha \|Bf\|^2 + \beta \|f\|^2 \rightarrow \min_f, \tag{15.13}$$

with $\alpha > 0$ and $\beta > 0$.

For the application of DRTLS and MPR one needs to select the operator B , and one needs a procedure, the so-called parameter choice rule [15], to select the appropriate parameters (α, β) . Parameter choice rules in the regularization methods need the noise levels in the considered ill-posed inverse problem [3]. In our case, as we mentioned in Sect. 15.2, we propose to view the operator $S_{\mathbf{x}}^* S_{\mathbf{x}}$ and the right-hand side $S_{\mathbf{x}}^* \mathbf{y}$ in (15.5) as the noisy versions of the operator $\mathcal{J}^* \mathcal{J}$ and the right-hand side $\mathcal{J}^* f_p$ in (15.3).

Thus, for the parameter choice rules, we need to estimate perturbation levels measured by

$$\|\mathcal{J}^* \mathcal{J} - S_{\tilde{x}}^* S_{\tilde{x}}\| \text{ and } \|\mathcal{J}^* f_p - S_{\tilde{x}}^* \mathbf{y}\|. \tag{15.14}$$

These estimations are derived in the next section.

15.4 Estimations of the Operator and Data Noise

In the analysis of the problems in the statistical learning theory, one often assumes (see, e.g., [10]) that there are constants $\{\Sigma, M\} \subset (0, +\infty)$ such that

$$\int_Y \left(\exp\left(\frac{|y - f^\dagger(x)|}{M}\right) - \frac{|y - f^\dagger(x)|}{M} - 1 \right) p_{y|x}(y|x) dy \leq \frac{\Sigma^2}{2M^2} \tag{15.15}$$

for almost all $x \in X$.

Now, we specify the structure of the subspace $\mathcal{H} \subset L^2(X, p_x)$. Since for the functions $f \in \mathcal{H}$ we are interested in their values $f(x)$ for $x \in X$, it is natural to require that the functionals $f(x)$ are continuous on \mathcal{H} . Reproducing Kernel Hilbert spaces (RKHS) [2, 6, 12] gives a rich variety of such spaces.

An RKHS is defined by a symmetric positive definite function $K(x, \tilde{x}) : X \times X \rightarrow \mathbb{R}$, which is called a kernel. Let us recall that a function $K(x, \tilde{x})$ is symmetric if $K(x, \tilde{x}) = K(\tilde{x}, x)$, and it is positive definite if for any $n \in \mathbb{N}$, any $\{x_1, \dots, x_n\} \subset X$, and any $\{a_1, \dots, a_n\} \subset \mathbb{R}$, with at least one $a_i \neq 0$,

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j K(x_i, x_j) > 0. \tag{15.16}$$

This property allows to define the scalar product for the functions of the form

$$\begin{aligned} f(x) &= \sum_{i=1}^n a_i K(x, x_i), \\ g(x) &= \sum_{j=1}^m b_j K(x, \tilde{x}_j) \end{aligned} \tag{15.17}$$

as follows:

$$(f, g)_K = \sum_{i=1}^n \sum_{j=1}^m a_i b_j K(x_i, \tilde{x}_j). \tag{15.18}$$

The RKHS that is defined (induced) by K is built as the completion of the space of all finite linear combinations (15.17) with respect to the norm that is induced by the scalar product (15.18). This RKHS is denoted by \mathcal{H}_K .

Let us also note the following property of linear combinations (15.17) that easily follows from (15.16).

Proposition 1. *Two functions $f(x)$ and $g(x)$ of the form*

$$f(x) = \sum_{i=1}^n a_i K(x, x_i), \{a_i\}_{i=1}^n \subset \mathbb{R},$$

$$g(x) = \sum_{i=1}^n \tilde{a}_i K(x, x_i), \{\tilde{a}_i\}_{i=1}^n \subset \mathbb{R}$$

are equal if and only if $a_i = \tilde{a}_i$ for $i = 1, \dots, n$.

It is common to put the following additional assumptions on the kernel K [4, 11].

Assumption 1. The kernel K is measurable. It is bounded with

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

The induced RKHS \mathcal{H}_K is separable.

With (15.15) and Assumption 1, we derive the estimates for the operator and data noise (15.14) in the following proposition.

Proposition 2. *Let f^\dagger be the solution of (15.3) with $\mathcal{H} = \mathcal{H}_K$, and let (15.15) and Assumption 1 hold. For $\eta \in (0, 1]$, consider the following set of events:*

$$G_\eta = \{\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in (X \times Y)^n \mid \|\mathcal{J}^* \mathcal{J} - S_{\mathbf{x}}^* S_{\mathbf{x}}\| \leq h, \|\mathcal{J}^* f_p - S_{\mathbf{x}}^* \mathbf{y}\| \leq \delta\},$$

with

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

$$\delta = \delta(n, \eta) = 2 \left(\frac{\kappa M}{n} + \frac{\kappa \Sigma + \sqrt{2}\kappa^2 \|f^\dagger\|}{\sqrt{n}} \right) \log \frac{4}{\eta}.$$

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

Proof. In [4], the following set of events was considered:

$$G'_\eta = \{\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in (X \times Y)^n \mid \|\mathcal{J}^* \mathcal{J} - S_{\mathbf{x}}^* S_{\mathbf{x}}\| \leq h, \|S_{\mathbf{x}}^* S_{\mathbf{x}} f^\dagger - S_{\mathbf{x}}^* \mathbf{y}\| \leq \delta'\},$$

with $\delta' = \delta'(n, \eta) = 2 \left(\frac{\kappa M}{n} + \frac{\kappa \Sigma}{\sqrt{n}} \right) \log \frac{4}{\eta}$. Using the results from [9, 14, 28], it was shown in [4] that $P[G'_\eta] \geq 1 - \eta$.

Now, consider $\mathbf{z} \in G'_\eta$, and let us estimate the corresponding data noise from (15.14):

$$\begin{aligned} \|\mathcal{J}^* f_p - S_{\mathbf{x}}^* \mathbf{y}\| &= \|\mathcal{J}^* \mathcal{J} f^\dagger - S_{\mathbf{x}}^* \mathbf{y}\| \leq \|\mathcal{J}^* \mathcal{J} f^\dagger - S_{\mathbf{x}}^* S_{\mathbf{x}} f^\dagger\| + \|S_{\mathbf{x}}^* S_{\mathbf{x}} f^\dagger - S_{\mathbf{x}}^* \mathbf{y}\| \\ &\leq h(n, \eta) \|f^\dagger\| + \delta'(n, \eta) = 2 \left(\frac{\kappa M}{n} + \frac{\kappa \Sigma + \sqrt{2} \kappa^2 \|f^\dagger\|}{\sqrt{n}} \right) \log \frac{4}{\eta}. \end{aligned}$$

Thus, $\mathbf{z} \in G_\eta$; therefore, $G_\eta \supset G'_\eta$, and

$$P[G_\eta] \geq P[G'_\eta] \geq 1 - \eta.$$

□

Remark 15.4.1. Since $\frac{1}{n} \leq \frac{1}{\sqrt{n}}$ for $n \in \mathbb{N}$, the considered errors can be estimated as

$$\begin{aligned} \|\mathcal{J}^* \mathcal{J} - S_{\mathbf{x}}^* S_{\mathbf{x}}\| &\leq \frac{c_h}{\sqrt{n}}, \\ \|\mathcal{J}^* f_p - S_{\mathbf{x}}^* \mathbf{y}\| &\leq \frac{c_\delta}{\sqrt{n}}, \end{aligned} \tag{15.19}$$

with some constants $\{c_h, c_\delta\} \subset (0, +\infty)$. These estimations can be used in the numerical realization of the regularization methods, which are used for solving (15.5).

15.5 Numerical Realization and Tests

In order to apply DRTLS and MPR to the operator equation (15.5) with $\mathcal{H} = \mathcal{H}_K$ one has to choose the weighted operator B . The simplest choice of this operator is the identity operator $I : \mathcal{H}_K \rightarrow \mathcal{H}_K$. With this choice, both DRTLS and MPR become the Tikhonov regularization (TR). Now, let us check the extrapolating properties of the estimators, which are obtained by TR.

In the context of the extrapolating estimators, additionally to the inputs $\{x_i\}_{i=1}^n$, which are presented in the given empirical data \mathbf{z} , one also deals with the inputs $\{x_i\}_{i=n+1}^m$ for which the corresponding outputs $\{y_i\}_{i=n+1}^m$ are not known. Moreover, the additional inputs $\{x_i\}_{i=n+1}^m$ are usually outside the $X_e := \text{conv}\{x_i, i = 1, \dots, n\}$. Thus, for a good extrapolating estimator, one expects additionally to a good approximation of the ideal estimator f^\dagger over the set X_e also a good approximation of f^\dagger over the $\text{conv}\{x_i, i = n + 1, \dots, m\}$.

In the statistical learning theory, the following function is often used as an ideal estimator for testing learning algorithms (e.g., [26]):

$$f^\dagger(x) = \frac{1}{10} \left(x + 2 \left(e^{-8\left(\frac{4\pi}{3}-x\right)^2} - e^{-8\left(\frac{\pi}{2}-x\right)^2} - e^{-8\left(\frac{3\pi}{2}-x\right)^2} \right) \right), \quad x \in [0, 2\pi].$$

This function belongs to the RKHS that is generated by the kernel $K(x, \tilde{x}) = x\tilde{x} + \exp(-8(x - \tilde{x})^2)$. We will use the RKHS that is generated by this kernel as the space \mathcal{H} .

The inputs \mathbf{x} in the empirical data are taken as follows:

$$x_i = \frac{\pi}{10}(i-1), \quad i = 1, \dots, 15, \quad (15.20)$$

and the outputs \mathbf{y} in the empirical data are generated as follows:

$$y_i = f^\dagger(x_i) + \hat{\delta}\xi_i, \quad i = 1, \dots, 15, \quad (15.21)$$

where $\{\xi_i\}$ are independent random variables with the uniform distribution over $[-1, 1]$. We consider $\hat{\delta} = 0.02$.

The estimator f_β that is constructed by TR with $\mathcal{H} = \mathcal{H}_K$ has the following representation:

$$f_\beta = \sum_{i=1}^n c_i K(x, x_i). \quad (15.22)$$

The coefficients $\mathbf{c} = (c_1, c_2, \dots, c_n)'$ in this representation satisfy the following system of linear equations (e.g., [20, 30]):

$$(\mathbf{K} + \beta n \mathbf{I})\mathbf{c} = \mathbf{y}, \quad (15.23)$$

where \mathbf{I} is the identity matrix of order n and $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n$.

Now, consider the situation when there is an additional input $x_{16} = \frac{\pi}{10}15$. Denote $\|f - g\|_{[a,b]}^\infty := \max_{x \in [a,b]} |f(x) - g(x)|$. In Fig. 15.1, one sees the estimator f_β , which is constructed by TR, and has the minimal extrapolating error $\min_{\beta \in (0,1]} \|f^\dagger - f_\beta\|_{[x_{15}, x_{16}]}^\infty$.

While it is possible to find such an estimator f_β that has a rather small interpolating error $\|f^\dagger - f_\beta\|_{[x_1, x_{15}]}^\infty$, the result in Fig. 15.1 shows that TR-estimators have rather bad extrapolating properties. Thus, other choices for the operator B in DRTLS and MPR are needed.

The sampling operator, which is scaled with the factor \sqrt{n} for convenience, i.e., $\sqrt{n}S_{\mathbf{x}} : \mathcal{H}_K \rightarrow \mathbb{R}^n$, can be proposed as a next choice for the operator B . Such an operator can be viewed as a statistical approximation of the identity operator. But in the contrast to the identity operator such a choice leads to a multiparameter regularization method that is different from TR. In this case, in the application of DRTLS to (15.5), one considers for several pairs of the parameters (α, β) the following operator equation:

$$(T_{\mathbf{x}}^* T_{\mathbf{x}} + \alpha n S_{\mathbf{x}}^* S_{\mathbf{x}} + \beta I)f = T_{\mathbf{x}}^* S_{\mathbf{x}}^* \mathbf{y}, \quad (15.24)$$

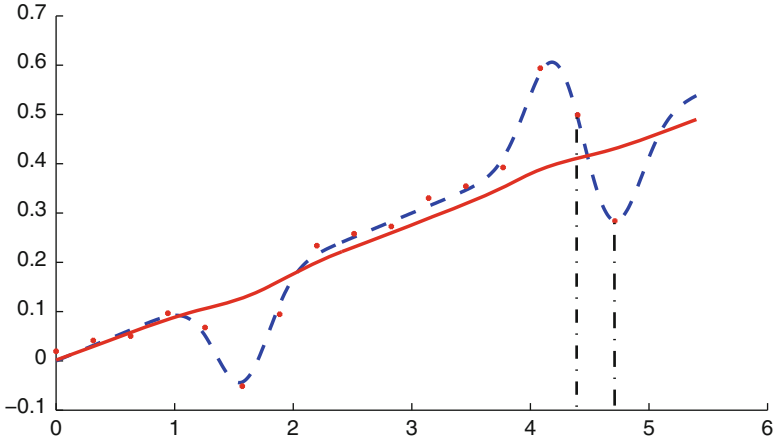


Fig. 15.1 The graph of the TR-estimator f_β (red curve) with the smallest extrapolating error $\|f^\dagger - f_\beta\|_{[x_{15}, x_{16}]}$. Red points correspond to the empirical data (15.20), (15.21). Blue dashed curve is the graph of the ideal estimator f^\dagger . The extrapolating interval $[x_{15}, x_{16}]$ is located between two vertical black dashed lines

where $T_x := S_x^* S_x$. As in the case of TR, the estimator $f_{\alpha, \beta}$ that is constructed by DRTLs with $\mathcal{H} = \mathcal{H}_K$ and $B = \sqrt{n} S_x$ has the representation (15.22). The coefficients \mathbf{c} in this representation satisfy the system of linear equations that is derived in the next proposition.

Proposition 3. *The function $f \in \mathcal{H}_K$ of the form*

$$f(x) = \sum_{i=1}^n c_i K(x, x_i) \tag{15.25}$$

solves the operator equation (15.24) with $\mathcal{H} = \mathcal{H}_K$ if and only if the coefficients $\mathbf{c} = (c_1, c_2, \dots, c_n)'$ satisfy the following system of linear equations:

$$[\mathbf{K}^2 + n^2(\alpha\mathbf{K} + \beta\mathbf{I})]\mathbf{c} = \mathbf{K}\mathbf{y}, \tag{15.26}$$

where \mathbf{I} is the identity matrix of order n and $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n$.

Proof. The derivation of the system (15.26) is similar to the derivation of the system (15.23) (see, e.g., [20, 30]).

It can be shown (e.g., [11, 20]) that the operator $S_x^* : \mathbb{R}^n \rightarrow \mathcal{H}_K$ is the following:

$$(S_x^* \mathbf{y})(x) = \frac{1}{n} \sum_{i=1}^n K(x, x_i) y_i.$$

For the functions of the form (15.25) we have that

$$T_{\mathbf{x}}f = S_{\mathbf{x}}^*S_{\mathbf{x}}f = \frac{1}{n} \sum_{i=1}^n K(x, x_i) \sum_{j=1}^n K(x_i, x_j)c_j.$$

Since $T_{\mathbf{x}}^* = T_{\mathbf{x}}$, we get that

$$T_{\mathbf{x}}^*S_{\mathbf{x}}^*y = T_{\mathbf{x}}S_{\mathbf{x}}^*y = \frac{1}{n^2} \sum_{i=1}^n K(x, x_i) \sum_{j=1}^n K(x_i, x_j)y_j,$$

$$T_{\mathbf{x}}^*T_{\mathbf{x}}f = T_{\mathbf{x}}^2f = \frac{1}{n^2} \sum_{i=1}^n K(x, x_i) \sum_{k=1}^n K(x_i, x_k) \sum_{j=1}^n K(x_k, x_j)c_j.$$

Thus, substituting the function (15.25) into the equation (15.24), we obtain in the left- and right-hand side of this equation a linear combination of functions $\{K(x, x_i)\}_{i=1}^n$. Since these linear combinations are equal only if their coefficients are equal (Proposition 1), we obtain the system of linear equations (15.26). \square

In the case when the additional inputs $\{x_i\}_{i=n+1}^m$ are given, it makes sense to include them into the sampling operator for the operator B . So, let us denote all given inputs as $\tilde{\mathbf{x}} = \{x_i\}_{i=1}^m$. Then, instead of $B = \sqrt{n}S_{\mathbf{x}}$, one can propose to consider

$$B = \sqrt{m}S_{\tilde{\mathbf{x}}}. \tag{15.27}$$

The estimator $f_{\alpha, \beta}$, which is constructed by DRTLS with such an operator B , has the following representation:

$$f(x) = \sum_{i=1}^m c_i K(x, x_i). \tag{15.28}$$

The system of linear equations for the coefficients \mathbf{c} can be derived similarly to the system (15.26). This system is the following:

$$[\mathbf{J}'\mathbf{K}\mathbf{J}\tilde{\mathbf{K}} + n^2(\alpha\tilde{\mathbf{K}} + \beta\tilde{\mathbf{I}})] \mathbf{c} = \mathbf{J}'\mathbf{K}\mathbf{y},$$

where $\tilde{\mathbf{I}}$ is the identity matrix of order m , $\tilde{\mathbf{K}} = (K(x_i, x_j))_{i, j=1}^m$, and $\mathbf{J} = (a_{ij} \mid i = 1, \dots, n; j = 1, \dots, m)$ with $a_{ii} = 1$, and $a_{ij} = 0$ when $i \neq j$.

Now, let us check the extrapolating properties of the estimators, which are constructed by DRTLS with B from (15.27). Let us take the empirical data from the test of TR, i.e., (15.20), (15.21), and let us consider two cases of additional inputs:

1. One additional input:

$$x_{16} = \frac{\pi}{10}15; \tag{15.29}$$

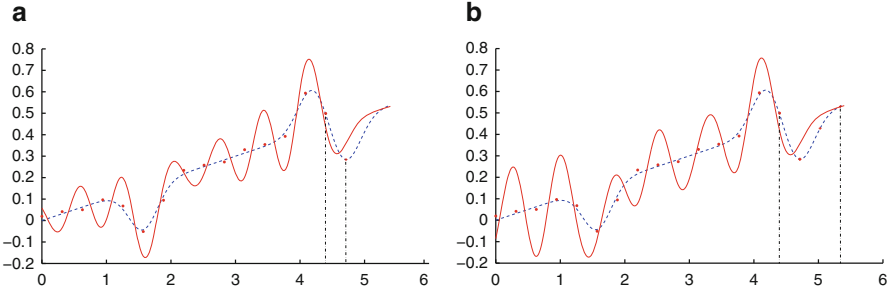


Fig. 15.2 The graphs of the DRTLs-estimators $f_{\alpha,\beta}$ (red curves) with the smallest extrapolating errors $\|f^\dagger - f_{\alpha,\beta}\|_{[x_{15},x_{16}]}^\infty$ (a) and $\|f^\dagger - f_{\alpha,\beta}\|_{[x_{15},x_{18}]}^\infty$ (b). Red points correspond to the empirical data (15.20), (15.21). Blue dashed curve is the graph of the ideal estimator f^\dagger . The extrapolating intervals $[x_{15},x_{16}]$ (a) and $[x_{15},x_{18}]$ (b) are located between two vertical black dashed lines

2. Three additional inputs:

$$x_i = \frac{\pi}{10}i, \quad i = 16, 17, 18. \tag{15.30}$$

In Fig. 15.2a, b, one sees the estimators $f_{\alpha,\beta}$, which are constructed by DRTLs and which have the minimal extrapolating errors. These estimators have better extrapolating properties than estimators that are constructed by TR, but the approximation of the ideal estimator over the set X_e is rather poor. Can another choice of the operator B improve this situation?

Recently [5], in the context of the statistical learning theory the following penalty functional was considered:

$$\rho(f) = \sum_{i,j=1}^m (f(x_i) - f(x_j))^2 w_{ij},$$

where w_{ij} are weights factors, which can be interpreted as edge weights in the data adjacency graph and are usually taken as $w_{ij} = \exp(-(x_i - x_j)^2)$. This functional can be represented as

$$\rho(f) = \|Bf\|_K^2, \quad \text{with } B = (S_{\tilde{x}}^* L S_{\tilde{x}})^{1/2}, \tag{15.31}$$

where the matrix L is the so-called graph Laplacian that is given by $L = D - W$, $W = (w_{ij})_{i,j=1}^m$, $D = (d_{ii})_{i,j=1}^m$ is a diagonal matrix with $d_{ii} = \sum_{j=1}^m w_{ij}$. Thus, $\rho(f)$ can be used in DRTLs.

As in the previous choice of the operator B , it can be shown that the estimator $f_{\alpha,\beta}$, which is constructed by DRTLs with B from (15.31), has the representation (15.28). The coefficients \mathbf{c} in this representation satisfy the following system of linear equations:

$$[\mathbf{J}'\mathbf{K}\mathbf{J}\tilde{\mathbf{K}} + n^2(\alpha L\tilde{\mathbf{K}} + \beta\tilde{\mathbf{I}})] \mathbf{c} = \mathbf{J}'\mathbf{K}\mathbf{y}.$$

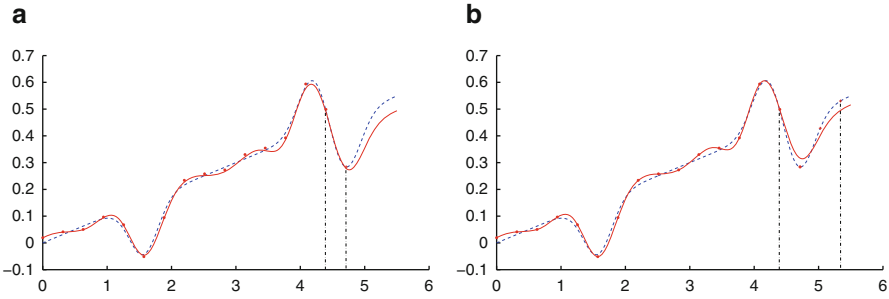


Fig. 15.3 The graphs of the MPR-estimators $f_{\alpha,\beta}$ (red curves) with the smallest extrapolating errors $\|f^\dagger - f_{\alpha,\beta}\|_{[x_{15},x_{16}]}^\infty$ (a) and $\|f^\dagger - f_{\alpha,\beta}\|_{[x_{15},x_{18}]}^\infty$ (b). Red points correspond to the empirical data (15.20), (15.21). Blue dashed curve is the graph of the ideal estimator f^\dagger . The extrapolating intervals $[x_{15},x_{16}]$ (a) and $[x_{15},x_{18}]$ (b) are located between two vertical black dashed lines

Our numerical experiments show that the obtained estimators are similar to the estimators that correspond to the choice (15.27). Thus, the choice (15.31) of the operator B does not improve the estimators that are constructed by DRTLS. However, MPR with the operator B from (15.31) gives much better estimators. Note, that in contrast to DRTLS, in MPR both regularization parameters α, β are positive.

In Fig. 15.3a, b, one sees the estimators $f_{\alpha,\beta}$, which are constructed by MPR and which have the minimal extrapolating errors. These estimators have not only the best extrapolating properties among the estimators that were considered so far, but they also approximate well the ideal estimator on the set X_e .

In practice, as any regularization method, MPR requires a rule for the choice of the involved regularization parameters. Such a rule is proposed in the next section.

15.6 The Choice of the Regularization Parameters in MPR

The so-called discrepancy principle (DP) (see, e.g., [15]) is a well-known choice rule for the parameters in the regularization methods. Let us consider the general framework of the Sect. 15.3. Denote $\{f_r\}$ the family of the regularized solutions of (15.9) that are constructed by a regularization method. Then, according to DP, one chooses f_r such that

$$\|A_h f_r - g_\delta\| = C\delta, C > 1. \tag{15.32}$$

There is a difficulty in using DP for the operator equation (15.5). Namely, a sharp estimate of the noise level δ is not available. Although Proposition 2 and Remark 15.4.1 give theoretical estimations of the noise level δ , in practice the choice of the involved constants there, in particular the constant c_δ in (15.19), is not clear. Moreover, the y -values in the empirical data have often the form (15.21), and

a good estimate of $\hat{\delta}$ can be assumed to be known. In this case, it seems reasonable instead of the condition (15.32), which in the case of the inverse problem (15.5) has the form

$$\|S_{\mathbf{x}}^* S_{\mathbf{x}} f_r - S_{\mathbf{x}}^* \mathbf{y}\| = C\delta, \tag{15.33}$$

to consider the following condition:

$$\|S_{\mathbf{x}} f_r - \mathbf{y}\| = \hat{C}\hat{\delta}. \tag{15.34}$$

Note, that the norms in the above conditions are connected through the following estimate:

$$\|S_{\mathbf{x}}^* S_{\mathbf{x}} f_r - S_{\mathbf{x}}^* \mathbf{y}\| \leq \|S_{\mathbf{x}}^*\| \cdot \|S_{\mathbf{x}} f_r - \mathbf{y}\|.$$

Thus, the control of the modified discrepancy $\|S_{\mathbf{x}} f_r - \mathbf{y}\|$ leads to the control of the original discrepancy $\|S_{\mathbf{x}}^* S_{\mathbf{x}} f_r - S_{\mathbf{x}}^* \mathbf{y}\|$. This may be used in the theoretical justification of the condition (15.34).

In MPR $f_r = f_{\alpha,\beta}$, and the condition (15.34), as well as the original condition (15.33), does not uniquely identify the pair of the regularization parameters (α, β) . The set of parameters that satisfy (15.34) can be called the discrepancy curve [22].

Among the pairs (α, β) on the discrepancy curve, one can look for the pair that defines the estimator with good extrapolating properties. For this purpose we propose to employ the so-called quasi-optimality principle [32]. The whole procedure for the choice of the appropriate pair of the regularization parameters (α, β) is presented below.

In the numerical realization of the regularization methods, the discrete sets of the regularization parameters in the form of the geometric sequence are frequently used. So, let us consider the following sequence for the parameters β :

$$\beta_k = \beta_0 q^k, \quad q > 1, \quad k = 0, 1, \dots, k_{\max}.$$

For each β_k , let us determine α_k for which the condition (15.34) is satisfied, i.e.,

$$\|S_{\mathbf{x}} f_{\alpha_k, \beta_k} - \mathbf{y}\| = \hat{C}\hat{\delta}. \tag{15.35}$$

This can be done using the so-called model function approach [19, 25, 34].

Now, let us define a closeness functional $d(f_{\alpha,\beta}, f_{\alpha',\beta'})$ that describes how close in some sense is the estimator $f_{\alpha,\beta}$ to the estimator $f_{\alpha',\beta'}$. For example, if $x_b \in X$ is an input point of interest, which can be an input without the corresponding output as (15.29), then $d(\cdot, \cdot)$ can be taken as follows:

$$d(f_{\alpha,\beta}, f_{\alpha',\beta'}) = |f_{\alpha,\beta}(x_b) - f_{\alpha',\beta'}(x_b)|. \tag{15.36}$$

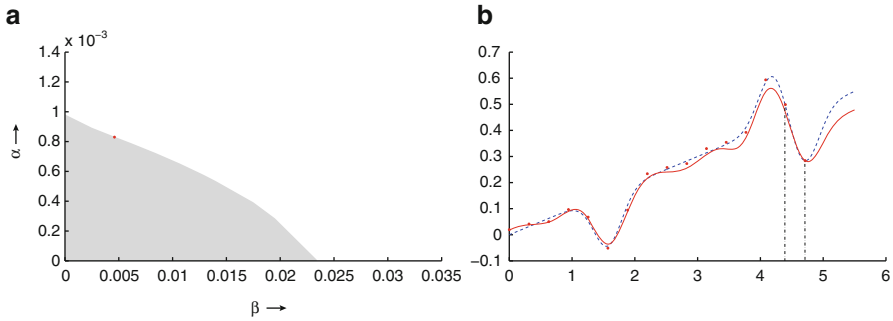


Fig. 15.4 (a) The discrepancy region with $\hat{C} = 1$. The red point corresponds to the pair (α_k, β_k) that is selected by the principle (15.37). (b) The graph of the corresponding MPR-estimator f_{α_k, β_k} . Red points correspond to the empirical data (15.20), (15.21). Blue dashed curve is the graph of the ideal estimator f^\dagger . The extrapolating interval $[x_{15}, x_{16}]$ is located between two vertical black dashed lines

Using the idea of the quasi-optimality principle and the chosen closeness functional d , among the pairs (α_k, β_k) that satisfy (15.35), one chooses such a pair that minimizes $d(f_{\alpha_k, \beta_k}, f_{\alpha_{k-1}, \beta_{k-1}})$, i.e., one chooses the pair (α_k, β_k) with the following index k :

$$k = \operatorname{argmin}_{k=1, \dots, k_{\max}} \{d(f_{\alpha_k, \beta_k}, f_{\alpha_{k-1}, \beta_{k-1}})\}. \tag{15.37}$$

Let us test the proposed procedure. Consider the empirical data (15.20) and (15.21). For these data $\hat{\delta} = 0.02$. Let us consider one additional input x_{16} from (15.29). First, let us take $\hat{C} = 1$. As the closeness functional d , we take (15.36) with $x_b = x_{16}$. In Fig. 15.4a, the discrepancy region, i.e., the region that contains (α, β) that satisfy

$$\|S_{\mathbf{x}} f_{\alpha, \beta} - \mathbf{y}\| \leq \hat{C} \hat{\delta}$$

is presented. The red point depicts the pair (α_k, β_k) that is selected by the principle (15.37). In Fig. 15.4b, the corresponding estimator f_{α_k, β_k} is presented. One observes that the chosen estimator is rather close to the best extrapolating estimator in Fig. 15.3a, which demonstrates effectiveness of the proposed parameters choice rule.

By varying the value of the constant \hat{C} one can obtain even better estimators. This is demonstrated in Fig. 15.5, where the results for $\hat{C} = 0.1$ can be found. This suggests that the influence of the constant \hat{C} should be studied in detail.

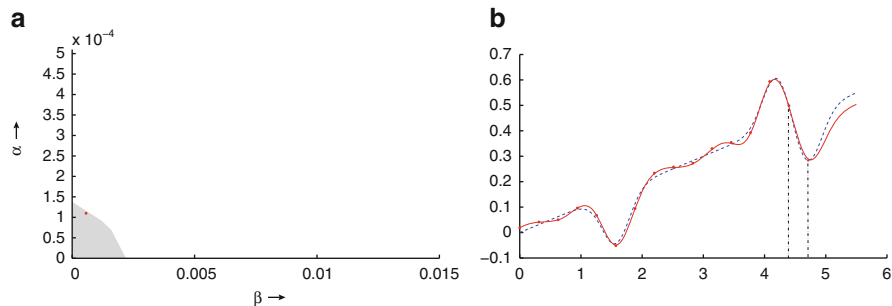


Fig. 15.5 (a) The discrepancy region with $\hat{C} = 0.1$. The red point corresponds to the pair (α_k, β_k) that is selected by the principle (15.37). (b) The graph of the corresponding MPR-estimator f_{α_k, β_k} . Red points correspond to the empirical data (15.20), (15.21). Blue dashed curve is the graph of the ideal estimator f^\dagger . The extrapolating interval $[x_{15}, x_{16}]$ is located between two vertical black dashed lines

15.7 Conclusions and Outlook

Construction of good extrapolating estimators requires novel approaches to the problem of constructing an estimator in the statistical learning theory, which can be formulated as an operator equation. We showed that this operator equation can be viewed as a perturbed operator equation with a perturbed operator and perturbed right-hand side. This view suggests the application of the multi-parameter regularization methods, such as DRTLS and MPR. Our numerical tests showed that among the considered realizations of DRTLS and MPR, a realization of MPR gives best extrapolating estimators, and thus, it can be proposed as a method of choice for constructing good extrapolating estimators. As any regularization method, MPR requires an automatic procedure for selecting the involved regularization parameters. We proposed such a procedure and demonstrated its successful performance.

Future research can be concentrated in the following directions.

We derived the perturbation levels in the operator equation for the estimator. This can be considered as a first step in the analysis of the application of multiparameter regularization methods, in particular MPR, for construction of extrapolating estimators. This analysis should be continued until the derivation of the estimates of the estimator general and extrapolating errors.

Other B -operators in MPR, such as (15.27), can be tried.

It is notable that with (15.27) the system of linear equations, which appears in the numerical realization of the corresponding MPR, has simpler structure than with (15.31). Thus, it is of particular interest to compare the quality of the extrapolating estimators that are constructed by these realizations of MPR.

One can also view

$$S_x f = y \tag{15.38}$$

as a perturbed operator equation. Application of DRTLS and MPR to (15.38) is quite straightforward, and it is remarkable that the systems of linear equations, which appear in the numerical realization, have a simpler structure in comparison to the systems that arise in the application of DRTLS and MPR to the operator equation (15.5). It remains to be verified whether this application leads to better extrapolating estimators. It should be also noted that the estimation and the interpretation of the perturbation levels in (15.38) have to be addressed.

Finally, a theoretical justification of the proposed choice rule for the regularization parameters in MPR is required. A more detailed study of the influence of the constant \hat{C} in (15.34) and of the connection between the conditions (15.34) and (15.33) should be also done.

Acknowledgments S. Lu is supported by the National Natural Science Foundation of China (No.11101093) and Shanghai Science and Technology Commission (No.11ZR1402800, No.11PJ1400800). S. Sampath is supported by EU-project “DIAdvisor” performed within 7th Framework Programme of EC.

References

1. Alpaydin E (2004) Introduction to machine learning (adaptive computation and machine learning). MIT Press
2. Aronszajn N (1950) Theory of reproducing kernels. *Trans Am Math Soc* 68:337–404
3. Bakushinskii AB (1984) Remarks on choosing regularization parameter using the quasi-optimality and ratio criterion. *USSR Comp Math Math Phys* 24:181–182
4. Bauer F, Pereverzev S, Rosasco L (2007) On regularization algorithms in learning theory. *J Complex* 23:52–72
5. Belkin M, Niyogi P, Sindhvani V (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J Mach Learning Res* 7:2399–2434
6. Berg C, Christensen JPR, Ressel P (1984) Harmonic analysis on semigroups. Theory of positive definite and related functions. Springer, New York
7. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin
8. Brezinski C, Redivo-Zaglia M, Rodriguez G, Seatzu S (2003) Multi-parameter regularization techniques for ill-conditioned linear systems. *Numer Math* 94:203–228
9. Caponnetto A, De Vito E (2005) Fast rates for regularized least-squares algorithm. In: CBCL Paper 248/AI Memo 2005-013. Massachusetts Institute of Technology, Cambridge, MA
10. Caponnetto A, De Vito E (2007) Optimal rates for the regularized least-squares algorithm. *Found Comput Math* 7:331–368
11. Carmeli C, De Vito E, Toigo A (2006) Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Anal Appl Singap* 4:377–408
12. Cucker F, Smale S (2002) On the mathematical foundations of learning. *Bull Am Math Soc New Ser* 39:1–49
13. De Vito E, Rosasco L, Caponnetto A (2006) Discretization error analysis for Tikhonov regularization. *Anal Appl Singap* 4:81–99
14. De Vito E, Rosasco L, Caponnetto A, De Giovannini U, Odone F (2005) Learning from examples as an inverse problem. *J Mach Learning Res* 6:883–904
15. Engl HW, Hanke M, Neubauer A (1996) Regularization of inverse problems. Kluwer Academic Publishers, Dordrecht

16. Girosi F, Poggio T (1990) Regularization algorithms for learning that are equivalent to multilayer networks. *Science* 247:978–982
17. Hofmann B (1999) *Mathematik inverser Probleme*. Teubner, Stuttgart
18. Kirsch A (1996) *An introduction to the mathematical theory of inverse problems*. Springer, Berlin
19. Kunisch K, Zou J (1998) Iterative choices of regularization parameters in linear inverse problems. *Inverse Probl* 14:1247–1264
20. Kurkova V (2010) Learning as an inverse problem in reproducing kernel hilbert spaces. Technical report, institute of computer science, academy of sciences of the Czech Republic
21. Lu S, Pereverzev SV (2011) Multi-parameter regularization and its numerical realization. *Numer Math* 118:1–31
22. Lu S, Pereverzev SV, Shao Y, Tautenhahn U (2010) Discrepancy curves for multi-parameter regularization. *J Inv Ill-Posed Probl* 18:655–676
23. Lu S, Pereverzev SV, Tautenhahn U (2008) Dual regularized total least squares and multi-parameter regularization. *Comp Meth Appl Math* 8:253–262
24. Lu S, Pereverzev SV, Tautenhahn U (2009) Regularized total least squares: computational aspects and error bounds. *SIAM J Matrix Anal Appl* 31:918–941
25. Lu S, Pereverzev SV, Tautenhahn U (2010) A model function method in regularized total least squares. *Appl Anal* 89:1693–1703
26. Micchelli CA, Pontil M (2005) Learning the kernel function via regularization. *J Mach Learning Res* 6:1099–1125
27. Mitchell TM (1997) *Machine learning*. McGraw Hill, New York
28. Pinelis I, Sakhanenko A (1986) Remarks on inequalities for large deviation probabilities. *Theor Probab Appl* 30:143–148
29. Poggio T, Girosi F (1990) Networks for approximation and learning. *Notices of AMS* 78:1481–1497
30. Poggio T, Smale S (2003) The mathematics of learning: dealing with data. *Notices Am Math Soc* 50(5):537–544
31. Rieder A (2003) *Keine Probleme mit inversen Problemen. Eine Einführung in ihre stabile Lösung*. Vieweg, Wiesbaden
32. Tikhonov AN, Glasko VB (1965) Use of the regularization method in non-linear problems. *USSR Comp Math Math Phys* 5:93–107
33. Vapnik VN (1998) *Statistical learning theory*. Wiley, NY
34. Xie J, Zou J (2002) An improved model function method for choosing regularization parameters in linear inverse problems. *Inverse Probl* 18:631–643