Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2008-062
CBCL-275

October 16, 2008

Adaptive Kernel Methods Using the Balancing Principle

Ernesto De Vito, Sergei Pereverzyev, and Lorenzo Rosasco

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

ADAPTIVE KERNEL METHODS USING THE BALANCING PRINCIPLE

E. DE VITO, S. PEREVERZYEV, L. ROSASCO

Abstract. The regularization parameter choice is a fundamental problem in supervised learning since the performance of most algorithms crucially depends on the choice of one or more of such parameters. In particular a main theoretical issue regards the amount of prior knowledge on the problem needed to suitably choose the regularization parameter and obtain learning rates. In this paper we present a strategy, the balancing principle, to choose the regularization parameter without knowledge of the regularity of the target function. Such a choice adaptively achieves the best error rate. Our main result applies to regularization algorithms in reproducing kernel Hilbert space with the square loss, though we also study how a similar principle can be used in other situations. As a straightforward corollary we can immediately derive adaptive parameter choice for various kernel methods recently studied. Numerical experiments with the proposed parameter choice rules are also presented.

1. Introduction

Most supervised learning algorithms depends on some tuning parameter, whose correct choice is crucial to ensure a good performance of the solution. Examples are the regularization parameter in regularized least squares regression [19] or the complexity of the hypothesis space in empirical risk minimization [40]. Theoretical analyses often show that the error incurred by the algorithm is sum of two terms, sample and approximation errors, having opposite behavior with respect to the tuning parameter [11], so that in this context a *natural* parameter choice is given by balancing the two error contributions. In many cases this choice provides optimal convergence rates in a mini-max setting [4, 8, 36, 7] and we refer to it as *best paremeter choice*.

However, this parameter choice raises conceptual and practical issues since estimates of the approximation error depends on some a priori knowledge of the problem which is usually not available. Indeed the so called *no free lunch theorem* shows that any data-independent parameter choice can not achieve the best convergence rate [21]. To overcome this problem, a data-driven choice is needed, ensuring that error rate of the solution to achieve the unknown optimal rate. In the statistical literature this problem is known as the problem of adaptive model selection [21, 15]. In regression model with fixed design, classical model selection schemes include Akaike criterion, BIC among the others (see [22] for references). In the setting of learning, where the design is random, some well known techniques for adaptive parameter choice are based on complexity regularization (see [16, 2, 21, 5] for general references and also [3, 25]), on data splitting- e.g. hold-out and cross-validation (see [16] and more recently [17, 39, 9]) and aggregation [38].

Date: October 16, 2008.

Based on the relation between learning theory and the theory of regularization in inverse problems – see [32, 40, 19, 35, 13] and references therein, in this paper we study a data driven method for a regularization parameter choice, namely *the balancing principle*, that has received a lot of attention in the theory inverse problems. Such a method is a development of an approach proposed by [24] in the context of Gaussian regression, that has been studied in the context of inverse problems in the paper by [20] and eventually developed in a series of papers (see [26] and references therein). Instances of Lepskii methods has been considered in statistical learning for aggregation of classifiers [38] and for empirical risk minimization algorithms [23]. Here we develop on the approach proposed in [20] which is very naturally while considering regularized kernel methods. We stress that the usual approaches to a posteriori parameter choice in inverse problems cannot be used directly in the context of learning since used methods are based on estimates of the stability of regularization methods measured in the space where the element of interest (regression or target function) should be recovered. Indeed, in the context of learning theory, typically, such estimates are measured with respect to the expected risk which depends on the unknown probability measure.

The method we propose is simple, it requires no data splitting, and achieves adaptively the best possible error rates (given by the a priori error bound). The proposed method allows us to easily derive adaptive parameter choices achieving optimal rates for several kernel methods [8, 36, 41, 4, 7] and we believe it might serve as a general way to obtain adaptive regularization schemes on kernel spaces. The plan of the paper follows. In Section 2 we give some background on the setting of supervised learning and discuss in some detail the problem of adaptive regularization parameter choice. In Section 3 we informally present and discuss our main results. In Section 4 we state and prove such results. We conclude in Section 5 with some numerical experiments.

2. Regularized Learning and Adaptive Parameter Choice

In this section after recalling a few basic concepts in supervised learning and fixing the notation we discuss the problem of adaptive regularization parameter choice that motivates the study in this paper.

2.1. Some Background on Supervised Learning.

We consider the problem of supervised learning as a multivariate function approximation problem from random samples [32, 40, 16, 11, 21].

Data Model. The data we are given is a *training set*,

$$\mathbf{z} = (\mathbf{x}, \mathbf{y}) = (x_1, y_1), \dots, (x_n, y_n),$$

where $x \in X \subset \mathbb{R}^d$ and $y \in Y \subseteq \mathbb{R}$ in regression or $y = \pm 1$ in classification. The model underlying the data is a fixed but unknown probability measure ρ on $X \times Y$ and $\mathbf{z}$ is identically and independently distributed according to ρ. Our goal is not to recover the whole probability ρ, but a target function $f_\rho : X \to Y$ minimizing the *expected risk*

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

where $\ell : Y \times \mathbb{R} \to \mathbb{R}^+$ is the *loss function*, for example the square loss.

Hypotheses Space. In practice, learning algorithms cannot work in the whole target space where the expected risk is defined and the search for a solution is

confined to a *hypotheses space* $\mathcal{H}$. Once the target space is fixed the best possible solution is the (so called) *best in the model* $f_{\mathcal{H}}$ such that

$$\mathcal{E}(f_{\mathcal{H}}) = \min_{f \in \mathcal{H}} \mathcal{E}(f).$$

Note that existence and uniqueness of the solution to the above problem typically requires conditions on $\mathcal{H}$ and ℓ. In the following we assume throughout that $f_{\mathcal{H}}$ exists. Examples of hypotheses spaces are splines, convex or linear combination of learners, piecewise linear or polynomial functions and a choice we will focus on are reproducing kernel Hilbert (RKH) spaces [1, 11].

Algorithms and Performance Measures. An algorithm can be seen as a map $\mathbf{z} \to f_{\mathbf{z}}$, that given a training set provides us with an *estimator* $f_{\mathbf{z}}$ of $f_{\mathcal{H}}$. If we want to measure the quality of such an estimate we have to decide on an approximation measure and use some probabilistic tool since $f_{\mathbf{z}}$ is a random variable. We will measure the approximation either with respect to the expected risk, or with respect to a norm $\|\cdot\|$ if the estimator and the target function belong to some normed hypotheses space. From the probabilistic point of view, a minimal, yet natural, requirement is *consistency*,

$$\lim_{n \to \infty} \mathbb{P}\left(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) > \varepsilon\right) = 0, \ \forall \varepsilon \in \mathbb{R}^+,$$

ensuring that the performance improves as we get more samples and eventually reaches the best possible error. More quantitative requirements concern convergence rate and give information on the performance for finite samples. In words, we try to estimate, with a given confidence, how far the error of our estimator is from the best possible error for fixed number of examples n. These latter results are usually expressed via tail inequalities of the form

$$\mathbb{P}\left(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) > \varepsilon(n, \eta)\right) \leq \eta, \tag{1}$$

where $0 < \eta \leq 1$, or equivalently

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \leq \varepsilon(n, \eta), \tag{2}$$

where the above inequality holds with probability at least $1 - \eta$. Similarly when the hypotheses space is a normed space we might also consider probabilistic bounds as measured by the corresponding norm, so that with probability at least $1 - \eta$

$$\|f_{\mathbf{z}} - f_{\mathcal{H}}\|_{\mathcal{H}} \leq \varepsilon(n, \eta). \tag{3}$$

For example when the hypotheses space is a RKH space, estimates in the RKH norm allows us to get estimates on various norms (see discussion in [36]).

A substantial difference between asymptotic and finite sample results is that the latter requires some prior assumption on the target function [16, 21, 15]. The impact of such a fact on the design of a fully data driven algorithm is somewhat at the the basis of the study in this paper. We discuss this point in details in the next section.

2.2. **Algorithms Depending on a Regularization Parameter.** In the previous section we considered an algorithm as a map $\mathbf{z} \to f_{\mathbf{z}}$, but in practice most algorithms can be seen as a two steps procedure. The first step defines a family of solutions depending on a real parameter

$$\mathbf{z} \to f_{\mathbf{z}}^{\lambda}, \ \lambda > 0,$$

whereas the second step determines how to choose such a parameter as a function of the training set and/or the number of training set points, i.e. $\lambda_n = \lambda(n, \mathbf{z})$. The final estimator $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$ is obtained only when both steps are defined.

A possible idea for choosing λ is based on the following observation. Suppose we have, in analogy to (3), a reliable probabilistic error estimate of the excess risk, i.e.

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda}) - \mathcal{E}(f_{\mathcal{H}}) \leq \varepsilon(\lambda, n, \eta) \tag{4}$$

for all $\lambda > 0$. Then we can simply take the parameter $\lambda_o(n)$ minimizing such an estimate. If the bound is tight such a choice can be shown to be optimal in a suitable sense (see Remark 1 below). The above reasoning hides an important conceptual and practical problem. Typically estimates like (4) are the sum of two competing terms, i.e.

$$\varepsilon(\lambda, n, \eta) = \mathcal{S}(n, \eta, \lambda) + \mathcal{A}(\lambda). \tag{5}$$

The nature of such terms is different:

- the term $\mathcal{S}(n, \eta, \lambda)$ is the so called sample error, it quantifies the error due to random sampling and is typically studied via concentration inequalities giving rise to an explicit bound, which does not depend on the unknown probability distribution. Usually $\mathcal{S}(n, \eta, \lambda)$ is a decreasing function both on the number of examples and on the regularization parameter λ.
- The term $\mathcal{A}(\lambda)$ is called approximation error, it does not depend on the data, but only on the unknown probability distribution. By a theoretical argument, one can always assume that it is an increasing function of λ, going to zero when λ goes to zero, but the rate strongly depends on $f_{\mathcal{H}}$.

If a bound like (4) is given, we can see that the best possible choice $\lambda_o(n)$ arises from the balancing of these two competing terms, namely from a sample-approximation (or bias-variance) trade-off. If both terms are known we can simply take the value of λ minimizing their sum or we can consider the value of λ making the contribution of the two terms equal (the crossing point in Figure 1). The two choices might in general different but if the sample and approximation errors depend polynomially on λ, they are equivalent in terms of learning rates, that is dependence on the number of samples n. In the following we consider this latter value as the best trade-off between sample and approximation error, that is the value λ_o solving

$$\mathcal{S}(n, \eta, \lambda) = \mathcal{A}(\lambda). \tag{6}$$

The corresponding error is, with probability at least $1 - \eta$

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda_o(n)}) - \mathcal{E}(f_{\mathcal{H}}) \leq 2\mathcal{S}(n, \eta, \lambda_o) = 2\mathcal{A}(\lambda_o). \tag{7}$$

Before developing further our reasoning let us give an example.

Example 1 (Regularized Least Squares)**.** *Consider the regularized least square algorithm*

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

where $\mathcal{H}$ is a RKH space and $\|\cdot\|_{\mathcal{H}}$ the corresponding norm. For some u such that $\int |u(x)|^2 d\rho_X(x) < \infty$, assume the target function to satisfy

$$f_{\mathcal{H}} = L_K^r u, \quad L_K f(x) = \int_X K(x, s) f(s) d\rho_X(x), \tag{8}$$

where ρ_X denotes the marginal probability of ρ on X and K is the kernel generating the RKH space which is assumed to be bounded. In this case one can prove [8, 36, 4, 7] *that the following bound holds with probability at least* $1-\eta$

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda}) - \mathcal{E}(f_{\mathcal{H}}) \leq C \log(4/\eta)(\frac{1}{\lambda n} + \lambda^{2r}),\ 1/2 < r \leq 1,$$

where $\mathcal{E}(f) = \int (y - f(x))^2 d\rho(x,y)$ *and* C *does not depend to* n, η, λ. *The best possible choice for* λ *and the corresponding rates are*

$$\lambda_o(n) = n^{-\frac{1}{2r+1}},\ O(n^{-\frac{2r}{2r+1}}),\ 1/2 < r \leq 1,$$

where the regularity of the target function is encoded in the index r.

The above discussion (and the example) shows how the parameter choice $\lambda_o(n)$ depends on the regularity properties of $f_{\mathcal{H}}$ that are usually not known. This observation motivates the interest into *adaptive parameter choices*. Namely, *we aim at defining a parameter choice independent to the prior assumption encoded in* $\mathcal{A}$, *but still achieving the best possible rate*. Clearly, the best achievable rate depends on $\mathcal{A}$, that is on the problem at hand, and still the hope is to design a data driven procedure to select a parameter which *adaptively* achieves such a rate. More formally we aim at finding some parameter choice $\lambda_+ = \lambda_+(n, \mathbf{z})$ such that

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda_+}) - \mathcal{E}(f_{\mathcal{H}}) \leq 2C\mathcal{A}(\lambda_o(n)) \tag{9}$$

for some positive constant C.

The subject of adaptive statistical estimation has received considerable attention in recent years. As we mentioned in the introduction, examples of methods proposed in the literature include: complexity regularization [16, 2, 21, 5] , which is highly related to the structural risk minimization principle, cross validation procedures [16, 17, 39, 9] and aggregation [38]. In the next section we describe a possible approach to the problem of adaptive parameter choice which is an instance of this latter method. In particular we develop on a formulation of a method originally due to Lepskii [24] that has become popular in the inverse problems literature where it is usually referred to as the *balancing principle* (see [26] and references therein). It is worth noting that variation of Lepskii type choices, also called pre-testing or comparison methods, have been previously considered in the context of statistical learning in the setting of aggregation of classifiers [38] and for empirical risk minimization algorithms [23]. The balancing principle as proposed in inverse problems is particularly natural while considering regularized learning algorithms.

We conclude this section with two remarks.

Remark 1 (Optimality and Minimax Results)**.** *In this paper we refer to the value* $\lambda_0(n)$, *defined by (6), as the best choice and to the corresponding rate as the best possible rate. However, the rate will be optimal in a minimax sense if the bound we started from (see (4), (5)) is tight. We do not discuss this problem and we refer to* [21, 15, 8] *for further information.*

Remark 2 (Optimality and Order Optimality)**.** *In our analysis we can usually compute the value of essentially all the constant appearing, but we do not expect such constants to be optimal. For this reason we often take fairly crude estimate and in fact we mainly focus on recovering the correct dependence on the number of samples. To some extent this is related to the difference between order optimality*

and optimality in inverse problems [18]. *Obtaining optimal values for the constants is clearly an interesting problem deserving further study.*

3. Adaptive Regularized Learning

In this section we informally describe the main results in the paper. As we previously mentioned, the parameter strategy we are going to discuss, namely the *balancing principle*, has become popular in the context of deterministic as well as stochastic inverse problems (see [20, 30, 27, 28] and references therein). In the following we start from this latter formulation and adapt it to the context of supervised learning.

Our main result deals with adaptive parameter selection for kernel methods when the error is measured via the excess risk, but we first present a preliminary result when the hypotheses space is a normed space and we measure the error via the norm in the space. These latter results can be of interest in their own and help appreciating the main intuition underlying the balancing principle.

3.1. **Adaptive Learning when the Error Measure is Known.** We assume both the estimator and the target to be elements of some normed space whose norm we denote with $\|\cdot\|$, so that we can consider

$$\|f_{\mathbf{z}} - f_{\mathcal{H}}\|^2 .$$

The important fact is that we assume such norm to be *known* (note that on the contrary $\mathcal{E}(\cdot)$ is not). Such a norm can be for example the norm $\|\cdot\|_{\mathcal{H}}$ in a (normed) hypotheses space or the empirical norm induced by the sample, that is

$$\|f\|^2_{\rho_{\mathbf{z}}} = \frac{1}{n}\sum_{i=1}^{n}(f(x_i))^2.$$

Again we assume that, for some regularization algorithm a bound of the form (5) is available, i.e. with probability $1-\eta$

$$\left\|f^{\lambda}_{\mathbf{z}} - f_{\mathcal{H}}\right\| \leq \mathcal{S}(n,\eta,\lambda) + \mathcal{A}(\lambda).$$

In the following we assume that the sample error is of the form

$$\mathcal{S}(n,\eta,\lambda) = \frac{\alpha(\eta)}{\omega(\lambda)\gamma(n)}$$

where $\alpha(\eta) > 1$. This latter assumption is typically satisfied and is made only to simplify the bounds and the exposition. For example, in the case of regularized least squares (see Example 2 below) $\omega(\lambda) = \sqrt{\lambda}$, $\gamma(n) = \sqrt{n}$ and $\alpha(\eta) = \log(4/\eta)$. Since $\alpha(\eta) > 1$, we can factorize the term $\alpha(\eta)$ and we have, with probability at least $1-\eta$, a bound of the form

$$\left\|f^{\lambda}_{\mathbf{z}} - f_{\mathcal{H}}\right\| \leq \alpha(\eta)\left(\frac{1}{\omega(\lambda)\gamma(n)} + \mathcal{A}(\lambda)\right), \tag{10}$$

where ω, $\mathcal{A}$ are assumed to be continuous, monotonically increasing functions and $\mathcal{A}(0) = 0$. The corresponding best parameter choice $\lambda_o(n)$ solves (6) and gives, with probability $1-\eta$, the rate

$$\left\|f^{\lambda_o(n)}_{\mathbf{z}} - f_{\mathcal{H}}\right\| \leq 2\alpha(\eta)\mathcal{A}(\lambda_o(n)).$$

To define a parameter strategy we first consider a suitable discretization for the possible values of the regularization parameter, that is an ordered sequence $(\lambda_i)_{i\in\mathbb{N}}$

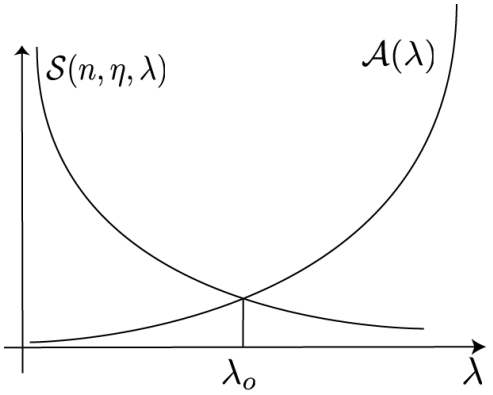

FIGURE 1. The figure represents the behavior of sample and approximation errors, respectively $\mathcal{S}(n,\eta,\lambda)$ and $\mathcal{A}(\lambda)$, as functions of λ, for fixed n,η.

such that the best value $\lambda_o(n)$ falls within the considered grid (see Section 4 for details). The balancing principle estimate for $\lambda_o(n)$ is defined via

$$\lambda_+ = \max\{\lambda_i : \left\| f_{\mathbf{z}}^{\lambda_i} - f_{\mathbf{z}}^{\lambda_j} \right\| \leq \frac{4\alpha(\eta)}{\omega(\lambda_j)\gamma(n)}, j = 0, 1, \ldots, i\}.$$

Such estimates no longer depend on $\mathcal{A}$ and the reason why we can expect it to be still sufficiently close to $\lambda_o(n)$ is better illustrated by Figure 1 and by the following reasoning.

Observe that if we take two values α, β such that $\alpha \leq \beta \leq \lambda_o(n)$ then with probability at least $1-\eta$

$$\begin{aligned} \left\| f_{\mathbf{z}}^{\alpha} - f_{\mathbf{z}}^{\beta} \right\| &\leq \left\| f_{\mathbf{z}}^{\alpha} - f_{\mathcal{H}} \right\| + \left\| f_{\mathbf{z}}^{\beta} - f_{\mathcal{H}} \right\| \\ &\leq \alpha(\eta) \left(\mathcal{A}(\alpha) + \frac{1}{\gamma(n)\omega(\alpha)} \right) + \\ &\quad \alpha(\eta) \left(\mathcal{A}(\beta) + \frac{1}{\gamma(n)\omega(\beta)} \right) \\ &\leq 4 \frac{\alpha(\eta)}{\gamma(n)\omega(\alpha)}. \end{aligned} \tag{11}$$

The intuition is that when such a condition is violated we are close to the intersection point of the two curves, that is to $\lambda_o(n)$. Such an intuition can be proved to be correct under mild assumptions. In fact, we will prove that, if

$$\omega(\lambda)\mathcal{A}(\lambda) \leq c\lambda, \tag{12}$$

for a suitable $c > 0$, and

$$\omega(\lambda_{i+1}) \leq q\omega(\lambda_i), \quad q > 1, \tag{13}$$

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

$$\left\| f_{\mathbf{z}}^{\lambda_+} - f_{\mathcal{H}} \right\| \leq 6q\alpha(\eta)\mathcal{A}(\lambda_o(n)).$$

The above parameter choice requires an extensive comparison of solutions at different values λ_i. The procedure can be simplified, at the price of slightly spoiling

the constant in the above inequality. In fact, we can take a geometric sequence $\lambda_i = \lambda_{\text{start}}\mu^i$, with $\mu > 1, \lambda_{\text{start}} \leq 1/(c\gamma(n))$ and introduce the choice

$$\bar{\lambda} = \max\{\lambda_i : \left\| f_{\mathbf{z}}^{\lambda_j} - f_{\mathbf{z}}^{\lambda_{j-1}} \right\| \leq \frac{4\alpha(\eta)}{\gamma(n)\omega(\lambda_{j-1})}, j = 0, 1, \ldots, i\},$$

requiring only comparison of solutions for adjacent parameter values. Again under mild assumptions one can prove that the following bound holds with probability at least $1 - \eta$

$$\left\| f_{\mathbf{z}}^{\bar{\lambda}} - f_{\mathcal{H}} \right\| \leq \alpha(\eta)\bar{C}\mathcal{A}(\lambda_o(n)),$$

where $\bar{C}$ does not depend on n and can be explicitly given.
We discuss some cases where the above results apply. The letter C is used to indicate constants independent to λ and n. We first go back to the RLS algorithm.

Example 2 (Regularized Least Squares)**.** *Error estimates for the RLS algorithm are known both for the expected risk (see Example 1) and the RKH norm* [8, 36, 4]*. In this latter case with probability* $1 - \eta$

$$\left\| f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}} \right\|_{\mathcal{H}} \leq C \log(4/\eta)(\frac{1}{\lambda\sqrt{n}} + \lambda^{r-1/2}),\ 1/2 < r \leq 3/2,$$

(under the same assumption of Example 1). It is straightforward to check that the above estimate satisfies the conditions allowing an application of the balancing principle to achieve optimal rates in an adaptive way.

Example 3 (Spectral Regularization)**.** *More generally the RLS algorithm can be seen as a special case of a large class of regularized kernel methods, namely spectral regularization, studied in* [4] *and including also L2-boosting* [6, 42] *and kernel principal component regression* [22, 34]*). All such algorithms can be written as*

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

where $\mathbf{K}_{ij} = K(x_i, x_j)$, $\alpha = (\alpha_1, \ldots, \alpha_n)$ *and* $g_\lambda(\sigma) \to \sigma^{-1}$ *as* $\lambda \to 0$ *(see* [14, 4, 7] *for details).*

The prior assumption (8) can be generalized to $f_{\mathcal{H}} = \phi(L_K)v, \|v\|_{\mathcal{H}} \leq R$ *(for a large class of functions* ϕ*, including* $\phi(\sigma) = \sigma^s, s > 0$*), where* L_K *is the integral operator in (8) restricted to* $\mathcal{H}$*. The following bound is proved in* [4]*, with probability at least* $1 - \eta$

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

for any[1] $\lambda \geq n^{-1/2}$*.*

Example 4 (Tikhonov Regularization with Convex Loss)**.** *The RLS algorithm can be generalized to*

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

Recall that if the loss function is convex and bounded, it is also locally Lipschitz continuous so that

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda}) - \mathcal{E}(f_{\mathcal{H}}) \leq L_\lambda \left\| f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}} \right\|_{\mathcal{H}},$$

[1] In [4] a slightly weaker condition is considered.

where the Lipschitz constant L_λ might depend on λ. The following bound is proved in [33] *(see also* [10] *for a more general setting), with probability at least* $1-\eta$

$$\left\|f_{\mathbf{z}}^{\lambda}-f_{\mathcal{H}}\right\|_{\mathcal{H}} \leq C\log(2/\eta)\left(\frac{L_\lambda}{\lambda\sqrt{n}}+\phi(\lambda)\right).$$

For a large number of loss functions (see [33]*) the constant L_λ can be explicitly computed and $\omega(\lambda)=L_\lambda/\lambda$, ϕ satisfy the assumptions required to apply the balancing principle.*

Example 5 (Elastic Net Regularization)**.** *The elastic-net algorithm proposed in* [43]*, is studied in* [12] *in the context of learning with an infinite dimensional overcomplete dictionary $(\psi_\gamma)_{\gamma\in\Gamma}$. In this case we let $\ell_2(\Gamma)$ be the space of $\beta=(\beta_\gamma)_{\gamma\in\Gamma}$ such that $\sum_{\gamma\in\Gamma}|\beta_\gamma|^2<\infty$ and look for the estimator β_n^λ minimizing*

$$\min_{\beta\in\ell_2(\Gamma)}\{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\sum_{\gamma\in\Gamma}\beta_\gamma\psi_\gamma(x_i)\right)^2+\lambda\left(\sum_{\gamma\in\Gamma}|\beta_\gamma|+\epsilon\sum_{\gamma\in\Gamma}\beta_\gamma^2\right)\}.$$

where $\epsilon,\lambda>0$. If we assume the target function to have an expansion $f_{\mathcal{H}}=\sum_{\gamma\in\Gamma}\beta_\gamma^*\psi_\gamma$ such that $\sum_{\gamma\in\Gamma}|\beta_\gamma^*|<\infty$, then, for $\lambda>1/\sqrt{n}$, it is possible to prove [12] *that with probability at least* $1-\eta$

$$\left\|\beta_n^\lambda-\beta^*\right\|_2 \leq C\log(4/\eta)\left(\frac{1}{\lambda\sqrt{n}}+\phi(\lambda)\right)$$

where $\|\cdot\|_2$ is the norm in $\ell_2(\Gamma)$. Again the above bound can be shown to satisfy the assumption needed to use the balancing principle.

3.2. **Adaptive Learning for the Expected Risks.** Our further goal is the adaptation with respect to the error as measured by the expected risk. Note that in this latter case there is no straightforward application of the balancing principle since it would require comparison of $\mathcal{E}(f_{\mathbf{z}}^{\lambda_i})-\mathcal{E}(f_{\mathbf{z}}^{\lambda_{i-1}})$ and hence a knowledge of the distribution ρ.

To deal with this situation we make two restrictions: 1) we consider regularization algorithms $f_{\mathbf{z}}^{\lambda}$ into a hypotheses space $\mathcal{H}$ which is a RKH space, 2) we consider regularization algorithms based on the square loss function, so that

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

Then we assume that error estimates for fixed λ are available both for the expected risk

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda})-\mathcal{E}(f_{\mathcal{H}}) \leq \alpha(\eta)^2(\frac{\lambda}{(\omega(\lambda)\gamma(n))^2}+\lambda\mathcal{A}(\lambda)^2), \tag{14}$$

and the RKH space norm

$$\left\|f_{\mathbf{z}}^{\lambda}-f_{\mathcal{H}}\right\|_{\mathcal{H}} \leq \alpha(\eta)(\frac{1}{\omega(\lambda)\gamma(n)}+\mathcal{A}(\lambda)), \tag{15}$$

where in this case we assume [2] that $\alpha(\eta)>\max\{\log(2/\eta)^{1/4},1\}$ and $\gamma(n)=\sqrt{n}$. The way we wrote the estimates is no coincidence since it corresponds to how the two error estimates are typically related (see for example [4, 36]) and we are going

[2]The constant in the bounds can be different but, for the sake of simplicity, we assume them to be equal to 1.

to assume such a relation to hold. Such an assumption is motivated by the relation between expected risk for the square loss and the RKH space norm, and we discuss it in Section 4.2.

Note that, because of this relation, the best parameter choice $\lambda_o(n)$ is the same in both cases and is given by the solution of (6) but the rates are different, in fact we have

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda_o(n)}) - \mathcal{E}(f_{\mathcal{H}}) \leq \alpha(\eta)^2 \lambda_o(n) \mathcal{A}(\lambda_o(n))^2 \tag{16}$$

for the expected risk and

$$\left\| f_{\mathbf{z}}^{\lambda_o(n)} - f_{\mathcal{H}} \right\|_{\mathcal{H}} \leq \alpha(\eta) \mathcal{A}(\lambda_o(n)) \tag{17}$$

for the RKH space norm. The fact that the best possible parameter choice is the same for both error measures is a promising indication. A possible idea would be to remember [1] that for the RKH space norm

$$|f(x)| \leq C_x \|f\|_{\mathcal{H}}, \ \forall x \in X, f \in \mathcal{H}$$

so that we can think of using the bound in the RKH space norm to bound the expected risk and use the balancing principle as presented above. Unfortunately in this way we are not going to match the best error rate for the expected risk, as can be seen comparing (16) and (17).

To achieve adaptation with respect to the expected risk we preliminary need a result (see Proposition 1) showing that if (14) and (15) hold for $\lambda \geq n^{-1/2}$ then with probability at least $1 - \eta$

$$\left\| f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}} \right\|_{\rho_{\mathbf{z}}}^2 \leq \hat{C}^2 \alpha(\eta)^2 \left(\frac{\lambda}{(\omega(\lambda)\gamma(n))^2} + \lambda \mathcal{A}(\lambda)^2 \right),$$

where we recall that

$$\|f - f_{\mathcal{H}}\|_{\rho_{\mathbf{z}}}^2 = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - f_{\mathcal{H}}(x_i))^2.$$

The above result shows that if we have error estimates for the expected risk and the norm in $\mathcal{H}$, we can also prove an error estimate for the empirical norm induced by the sample.

Now, both the empirical norm and the RKH space norm are known so that we can use the balancing principle to define

$$\lambda_{\rho_{\mathbf{z}}} = \max\{\lambda_i : \left\| f_{\mathbf{z}}^{\lambda_i} - f_{\mathbf{z}}^{\lambda_j} \right\|_{\rho_{\mathbf{z}}}^2 \leq \frac{4\hat{C}\alpha(\eta)\sqrt{\lambda_j}}{\gamma(n)\omega(\lambda_j)}, j = 0, 1, \ldots, i\},$$

and

$$\lambda_{\mathcal{H}} = \max\{\lambda_i : \left\| f_{\mathbf{z}}^{\lambda_i} - f_{\mathbf{z}}^{\lambda_j} \right\|_{\mathcal{H}} \leq \frac{4\alpha(\eta)}{\gamma(n)\omega(\lambda_j)}, j = 0, 1, \ldots, i\},$$

that we know are going to achieve the best rates in the corresponding norms by a direct application of the balancing principle. Our main result shows that the choice

$$\hat{\lambda} = \min\{\lambda_{\rho_{\mathbf{z}}}, \lambda_{\mathcal{H}}\} \tag{18}$$

allows us to achieve best error rates for the expected risk in an adaptive way. In fact we will prove that if $\lambda\omega(\lambda)$, $\lambda\mathcal{A}(\lambda)$ are monotonically increasing functions with $\mathcal{A}(0) = 0$ and (12), (13) hold, then with probability at least $1 - \eta$

$$\mathcal{E}(f_{\mathbf{z}}^{\hat{\lambda}}) - \mathcal{E}(f_{\mathcal{H}}) \leq 6qC\alpha(\eta)^2 \lambda_o(n) \mathcal{A}(\lambda_o(n))^2$$

where the value of C can be explicitly given. As an application of the above result we show how it allows an optimal adaptive parameter choice for the class of spectral regularization algorithms studied in [14, 4]. First we illustrate the application to regularized least square algorithm.

Example 6 (Regularized Least Squares)**.** *As we previously mentioned, for regularized least square algorithm (see Example 1) we have with probability at least* $1-\eta$

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda}) - \mathcal{E}(f_{\mathcal{H}}) \leq C \log(4/\eta)(\frac{1}{\lambda n} + \lambda^{2r}),\ 1/2 < r \leq 1,$$

but also

$$\left\|f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}\right\|_{\mathcal{H}} \leq C \log(4/\eta)(\frac{1}{\lambda\sqrt{n}} + \lambda^{r-1/2}),\ 1/2 < r \leq 3/2.$$

Applying the above result we have that the parameter choice (18) satisfies with probability at least $1-\eta$

$$\mathcal{E}(f_{\mathbf{z}}^{\hat{\lambda}}) - \mathcal{E}(f_{\mathcal{H}}) \leq 6qC \log(4/\eta) n^{-\frac{2r}{2r+1}},\ 1/2 < r \leq 1.$$

Example 7 (Spectral Regularization)**.** *In Example 3 we have seen that RLS is a particular instance of a class of spectral algorithms for supervised learning. For this latter class of methods the following bound on the expected risk is known* [4] *to hold with probability at least* $1-\eta$

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda}) - \mathcal{E}(f_{\mathcal{H}}) \leq C \log(4/\eta)(\frac{1}{\lambda n} + \lambda\phi(\lambda)^2),$$

where ϕ *is a function encoding the smoothness of the target function (see Example 3 and* [4] *for details). Again it is easy to see that the assumptions to apply the balancing principle hold.*

We end this section with the following remark.

Remark 3 (Computing Balancing Principle)**.** *The proposed parameter choices can be computed exploiting the properties of RKH spaces. In fact for* $f = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot)$ *we have*

$$\begin{aligned} \|f\|_{\mathcal{H}}^2 &= \left\langle \sum_{i=1}^{n} \alpha_i K(x_i,\cdot), \sum_{i=1}^{n} \alpha_i K(x_i,\cdot) \right\rangle_{\mathcal{H}} \\ &= \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) = \alpha \mathbf{K} \alpha, \end{aligned}$$

where we used the reproducing property $\langle K(x,\cdot), K(s,\cdot)\rangle_{\mathcal{H}} = K(x,s)$ [1]*. Then we can check that for* $f_{\mathbf{z}}^{\beta} = \sum_{i=1}^{n} \alpha_i^{\beta} K(x_i,\cdot)$, $f_{\mathbf{z}}^{\lambda} = \sum_{i=1}^{n} \alpha_i^{\lambda} K(x_i,\cdot)$ *we have*

$$\begin{aligned} \left\|f_{\mathbf{z}}^{\beta} - f_{\mathbf{z}}^{\lambda}\right\|_{\mathcal{H}}^2 &= \alpha^{\beta}\mathbf{K}\alpha^{\beta} - 2\alpha^{\beta}\mathbf{K}\alpha^{\lambda} + \alpha^{\lambda}\mathbf{K}\alpha^{\lambda} \\ &= (\alpha^{\beta} - \alpha^{\lambda})\mathbf{K}(\alpha^{\beta} - \alpha^{\lambda}). \end{aligned}$$

Similarly one can see that

$$\left\|f_{\mathbf{z}}^{\beta} - f_{\mathbf{z}}^{\lambda}\right\|_{\rho_{\mathbf{z}}}^2 = (\alpha^{\beta} - \alpha^{\lambda})\mathbf{K}^2(\alpha^{\beta} - \alpha^{\lambda}).$$

4. Mathematical Results

In this section we give the proofs of the results we previously presented. Following the discussion of previous section we first prove the results when the error is measured with respect to some known norm.

4.1. **Results for Known Norm.** Our main assumptions regard the error estimate for fixed λ and specify suitable conditions on $\mathcal{A}$ and ω.

Assumption 1. *For $\lambda > 0$ both $f_{\mathbf{z}}^{\lambda}$ and $f_{\mathcal{H}}$ belong to some normed space and moreover with probability at least $1-\eta$*

$$\left\|f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}\right\| \leq \alpha(\eta)\left(\frac{1}{\omega(\lambda)\gamma(n)} + \mathcal{A}(\lambda)\right)$$

where:

- *$\omega(\lambda)$ is a continuous, increasing function;*
- *$\mathcal{A}(\lambda)$ is a continuous, increasing function with $\mathcal{A}(0) = 0$;*
- *$\omega(\lambda)\mathcal{A}(\lambda) \leq c\lambda$,*

and $\alpha(\eta) > 1$. Moreover, assume that the bound holds uniformly with respect to λ, meaning that the collection of training sets for which it holds with confidence $1-\eta$ does not depend on λ.

Recall that in this case the best parameter choice, solving (6), achieves the error estimate (9) and it can be shown that the last condition in Assumption 1 ensures $\lambda_o(n) \geq 1/(c\gamma(n))$. Note that, if we now restrict out attention to some discrete sequence $\lambda_{\text{start}} \leq 1/(c\gamma(n))$, then it is easy to see that the best estimate for $\lambda_o(n)$ is

$$\lambda_* = \max\{\lambda_i | \mathcal{A}(\lambda_i) \leq \frac{1}{\omega(\lambda_i)\gamma(n)}\}$$

which still depends on $\mathcal{A}$. Finally recall that by the balancing principle we select $\lambda_+ = \lambda_+(n, \mathbf{z})$ by

$$\lambda_+ = \max\{\lambda_i : \left\|f_{\mathbf{z}}^{\lambda_i} - f_{\mathbf{z}}^{\lambda_j}\right\| \leq \frac{4\alpha(\eta)}{\omega(\lambda_j)\gamma(n)}, j = 0, 1, \ldots, i\}$$

The following theorem shows that the choice λ_+ provides the same error estimate of $\lambda_o(n)$ up to a constant factor.

Theorem 1. *If Assumption 1 holds and moreover, for $\lambda_{start} \leq 1/(c\gamma(n))$, we have*

$$\omega(\lambda_{i+1}) \leq q\omega(\lambda_i), \ q > 1, \tag{19}$$

then with probability at least $1-\eta$

$$\left\|f_{\mathbf{z}}^{\lambda_+} - f_{\mathcal{H}}\right\| \leq 6q\alpha(\eta)\mathcal{A}(\lambda_o(n)).$$

Proof. Note that all the inequalities in the proof are to be intended as holding with probability at least $1-\eta$. Recall that by (11) for λ, β such that $\lambda \leq \beta \leq \lambda_o(n)$ we have

$$\left\|f_{\mathbf{z}}^{\lambda} - f_{\mathbf{z}}^{\beta}\right\| \leq \frac{4\alpha(\eta)}{\omega(\lambda)\gamma(n)}.$$

It is easy to prove that $\lambda_* \leq \lambda_+$. Indeed by definition $\lambda_* \leq \lambda_o(n)$, and we know that, for any $\lambda_j \leq \lambda_* \leq \lambda_o(n)$,

$$\left\|f_{\mathbf{z}}^{\lambda_j} - f_{\mathbf{z}}^{\lambda_*}\right\| \leq \frac{4\alpha(\eta)}{\omega(\lambda_j)\gamma(n)},$$

so that, in particular, $\lambda_* \leq \lambda_+$. From the definition of λ_+ and λ_* we get

$$\begin{aligned} \left\|f_{\mathbf{z}}^{\lambda_+} - f_{\mathcal{H}}\right\| &\leq \left\|f_{\mathbf{z}}^{\lambda_+} - f_{\mathbf{z}}^{\lambda_*}\right\| + \left\|f_{\mathbf{z}}^{\lambda_*} - f_{\mathcal{H}}\right\| \\ &\leq \frac{4\alpha(\eta)}{\omega(\lambda_*)\gamma(n)} + \alpha(\eta)\left(\mathcal{A}(\lambda_*) + \frac{1}{\omega(\lambda_*)\gamma(n)}\right) \\ &\leq \frac{4\alpha(\eta)}{\omega(\lambda_*)\gamma(n)} + \frac{2\alpha(\eta)}{\omega(\lambda_*)\gamma(n)} \leq \frac{6\alpha(\eta)}{\omega(\lambda_*)\gamma(n)}. \end{aligned} \tag{20}$$

Finally to relate λ_* and $\lambda_o(n)$, we let $\lambda_* = \lambda_\ell$ so that $\lambda_* = \lambda_\ell \leq \lambda_o(n) \leq \lambda_{\ell+1}$. Since $\omega(\lambda)$ is increasing, we can use (19) to get $\omega(\lambda_o(n)) \leq \omega(\lambda_{\ell+1}) \leq q\omega(\lambda_\ell) = q\omega(\lambda_*)$. The above reasoning yields

$$\frac{1}{\omega(\lambda_*)} \leq \frac{q}{\omega(\lambda_o(n))} \tag{21}$$

and if we plug the above inequality into (20), the definition of $\lambda_o(n)$ gives

$$\left\|f_{\mathbf{z}}^{\lambda_+} - f_{\mathcal{H}}\right\| \leq 6q\alpha(\eta)\frac{1}{\omega(\lambda_o(n))\gamma(n)} = 6q\alpha(\eta)\mathcal{A}(\lambda_o(n))$$

so that the theorem is proved. □

We now consider the case when the sequence of values for the regularization parameter is defined by a geometric sequence $\lambda_i = \lambda_{\text{start}}\mu^i$, with $\mu > 1, \lambda_{\text{start}} \leq 1/(c\gamma(n))$ and consider the parameter choice $\bar{\lambda} = \bar{\lambda}(n, \mathbf{z})$ defined by

$$\bar{\lambda} = \max\{\lambda_i : \left\|f_{\mathbf{z}}^{\lambda_j} - f_{\mathbf{z}}^{\lambda_{j-1}}\right\| \leq \frac{4\alpha(\eta)}{\omega(\lambda_{j-1})\gamma(n)}, j = 0, 1, \ldots, i\}.$$

Next theorem studies the error estimate obtained with such a choice.

Theorem 2. *If Assumption 1 holds and moreover, there are $b > a > 1$ such that for any $\lambda > 0$,*

$$\omega(2\lambda)/b \leq \omega(\lambda) \leq \omega(2\lambda)/a, \tag{22}$$

then with probability at least $1 - \eta$

$$\left\|f_{\mathbf{z}}^{\bar{\lambda}} - f_{\mathcal{H}}\right\| \leq C\alpha(\eta)\mathcal{A}(\lambda_o(n))$$

where C might depend on a, b, μ.

Proof. The proof follows exactly the one for deterministic inverse problems though inequalities here are to be intended with probability at least $1-\eta$. The key observation is that we can easily control the distance between the solutions corresponding to λ_* and $\bar{\lambda}$. In fact if we let $\lambda_* = \lambda_\ell$ and $\bar{\lambda} = \lambda_m$ clearly $m \geq \ell$ and we can use the definition of $\bar{\lambda}$ to write

$$\begin{aligned} \left\|f_{\mathbf{z}}^{\bar{\lambda}} - f_{\mathbf{z}}^{\lambda_*}\right\| &\leq \sum_{j=\ell+1}^{m} \left\|f_{\mathbf{z}}^{\lambda_j} - f_{\mathbf{z}}^{\lambda_{j-1}}\right\| \\ &\leq 4\alpha(\eta)\frac{1}{\gamma(n)} \sum_{j=\ell+1}^{m} \frac{1}{\omega(\lambda_{j-1})} \\ &\leq 4\alpha(\eta)\frac{1}{\gamma(n)} \sum_{j=0}^{m-\ell-1} \frac{1}{\omega(\lambda_*\mu^j)}. \end{aligned} \tag{23}$$

Now for any $\mu > 1$, $\alpha > 1$ let $p, s \in \mathbb{N}$ be such that $2^p \leq \mu \leq 2^{p+1}$ and $2^s \leq \alpha \leq 2^{s+1}$. Then using (22) we get

$$\begin{aligned} \frac{1}{\omega(\alpha\lambda_*)} &\leq \frac{1}{\omega(2^s\lambda_*)} \leq \frac{1}{a^s\omega(\lambda_*)} \leq \frac{1}{a^{\log_2 \alpha}\omega(\lambda_*)} \\ \omega(\lambda_i) &= \omega(\mu\lambda_{i-1}) \leq b^{p+1}\omega(\lambda_{i-1}) \leq b^{\log_2 2\mu}\omega(\lambda_{i-1}). \end{aligned}$$

The last inequality shows that (19) is satisfied with $q = b^{\log_2 2\mu}$ and also

$$\sum_{j=0}^{m-\ell-1} \frac{1}{\omega(\lambda_*\mu^j)} \leq \frac{1}{\omega(\lambda_*)} \frac{a^{\log_2 2\mu}}{a^{\log_2 \mu} - 1}.$$

Finally we can use the above inequality and the definition of λ_* to get

$$\begin{aligned} \left\| f_{\mathbf{z}}^{\bar\lambda} - f_{\mathcal{H}} \right\| &\leq \left\| f_{\mathbf{z}}^{\lambda_*} - f_{\mathcal{H}} \right\| + \left\| f_{\mathbf{z}}^{\bar\lambda} - f_{\mathbf{z}}^{\lambda_*} \right\| \\ &\leq 2\alpha(\eta)\frac{1}{\gamma(n)\omega(\lambda_*)} + 4\alpha(\eta)\frac{a^{\log_2 2\mu}}{a^{\log_2 \mu} - 1}\frac{1}{\gamma(n)\omega(\lambda_*)} \\ &\leq 2\alpha(\eta)\left(1 + 2\frac{a^{\log_2 2\mu}}{a^{\log_2 \mu} - 1}\right)\frac{b^{\log_2 2\mu}}{\gamma(n)\omega(\lambda_o(n))}. \end{aligned}$$

The theorem is proved recalling the definition of $\lambda_o(n)$. □

4.2. **Results for the expected risk.** In this section we prove the main result of the paper allowing adaptive regularization for kernel based algorithms. We assume the space X to be a separable metric space and consider a RKH space such that the corresponding reproducing kernel $K : X \times X \to \mathbb{R}$ is measurable and bounded, that is

$$\kappa = \sup_{x \in X} \sqrt{K(x, x)}. \tag{24}$$

We denote with ρ_X the marginal probability on X of the distribution ρ and with $\rho(y|x)$ the conditional probability. As we previously mentioned we restrict our attention to the square loss function so that

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

If $\int y^2\rho(x, y) < \infty$ the expected risk is a well defined functional on the space $L^2(X, \rho_X)$ of square integrable functions that is $f : X \to \mathbb{R}$ such that

$$\|f\|_\rho^2 = \int_X f(x)^2 d\rho_X(x) < \infty. \tag{25}$$

In this case some facts are well known (see for example [11, 21]). The minimizer of $\mathcal{E}(f)$ over $L^2(X, \rho_X)$ is the regression function

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

and for $f \in L^2(X, \rho_X)$ we can write

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

In other words the norm given by (25) provides a natural way to measure the approximation since it puts weights on points that are more likely to be sampled.

Then, it is clear that the application of the balancing principle is not straightforward since we should evaluate

$$\left\| f_{\mathbf{z}}^{\beta} - f_{\mathbf{z}}^{\lambda} \right\|_{\rho}.$$

The main result of this section shows that when the estimator belongs to a RKH space we can still define a data driven parameter choice achieving the best possible error estimate also in this case. The following concentration result will be crucial.

Proposition 1. *Assume that $\mathcal{H}$ is a RKH space with bounded kernel (24). For $f \in \mathcal{H}$ we have with probability at least $1-\eta$*

$$| \|f\|_{\rho} - \|f\|_{\rho_{\mathbf{z}}} | \leq C_{\kappa} (\frac{\log(2/\eta)}{n})^{1/4} \|f\|_{\mathcal{H}},$$

where

$$\|f\|_{\rho_{\mathbf{z}}}^2 = \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2$$

is the empirical norm and $C_{\kappa}^2 = 2\sqrt{2}\kappa^2$.

Proof. Let $K_x = K(x, \cdot)$, if $f \in \mathcal{H}$, by the reproducing property we have $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$. Then we can write

$$\begin{aligned} \|f\|_{\rho}^2 &= \int_X \langle f, K_x \rangle_{\mathcal{H}} \langle f, K_x \rangle_{\mathcal{H}} \, d\rho_X(x) \\ &= \left\langle f, \int_X \langle f, K_x \rangle_{\mathcal{H}} K_x d\rho_X(x) \right\rangle_{\mathcal{H}} =: \langle f, Tf \rangle_{\mathcal{H}}. \end{aligned}$$

Reasoning in the same way we get

$$\begin{aligned} \|f\|_{\rho_{\mathbf{z}}}^2 &= \frac{1}{n} \sum_{i=1}^{n} \langle f, K_{x_i} \rangle_{\mathcal{H}} \langle f, K_{x_i} \rangle_{\mathcal{H}} \\ &= \left\langle f, \frac{1}{n} \sum_{i=1}^{n} \langle f, K_{x_i} \rangle_{\mathcal{H}} K_{x_i} \right\rangle_{\mathcal{H}} =: \langle f, T_{\mathbf{x}} f \rangle_{\mathcal{H}}. \end{aligned}$$

The operators $T, T_{\mathbf{x}}$ can be shown to be positive and of Hilbert-Schmidt type [8]. From the above reasoning it follows that $\forall f \in \mathcal{H}$

$$| \|f\|_{\rho} - \|f\|_{\rho_{\mathbf{z}}} | \leq \sqrt{\|T - T_{\mathbf{x}}\|} \, \|f\|_{\mathcal{H}}. \tag{26}$$

The quantity $\|T - T_{\mathbf{x}}\|$ have been studied in [8, 4] and we just sketch how to deal with it. Since T, $T_{\mathbf{x}}$ and $\langle \cdot, K_x \rangle_{\mathcal{H}} K_x$ are Hilbert Schmidt operators then

$$\|T - T_{\mathbf{x}}\| \leq \|T - T_{\mathbf{x}}\|_{HS}.$$

The random variable $\xi : X \to HS(\mathcal{H})$, from the input space to the space of Hilbert-Schmidt operators, defined by

$$\xi = \langle \cdot, K_x \rangle_{\mathcal{H}} K_x - T$$

is a Hilbert space valued random variable with zero mean, since $T = \mathbb{E}\left(\langle \cdot, K_x \rangle_{\mathcal{H}} K_x\right)$, and bounded by $2\kappa^2$. Concentration inequalities for Hilbert space valued random variable [31] immediately yields with probability at least $1-\eta$

$$\|T - T_{\mathbf{x}}\| \leq \|T - T_{\mathbf{x}}\|_{HS} \leq \frac{(\log(2/\eta))^{1/2} C_{\kappa}^2}{\sqrt{n}}.$$

The theorem is proved plugging the above estimate into (26). □

We are now in position to state our main result. The following assumption is the analogous to Assumption 1.

Assumption 2. *Let $\mathcal{H}$ be a RKH space and assume it exists $f_{\mathcal{H}}$ s.t.*

$$\mathcal{E}(f_{\mathcal{H}}) = \min_{f\in\mathcal{H}} \mathcal{E}(f).$$

Assume that $f_{\mathbf{z}}^{\lambda}$ belongs to $\mathcal{H}$ and for $\lambda \geq n^{-1/2}$ the following bounds hold with probability at least $1-\eta$

$$\left\|f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}\right\|_{\rho} \leq \alpha(\eta)\sqrt{\lambda}\left(\frac{1}{\sqrt{n}\omega(\lambda)} + \mathcal{A}(\lambda)\right) \tag{27}$$

and

$$\left\|f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}\right\|_{\mathcal{H}} \leq \alpha(\eta)\left(\frac{1}{\sqrt{n}\omega(\lambda)} + \mathcal{A}(\lambda)\right), \tag{28}$$

where

- *$\sqrt{\lambda}\omega(\lambda)$ is a continuous, increasing function;*
- *$\sqrt{\lambda}\mathcal{A}(\lambda)$ is a continuous, increasing function with $\mathcal{A}(0) = 0$,*
- *$\omega(\lambda)\mathcal{A}(\lambda) \leq c\lambda$,*

and $\alpha(\eta) > \max\{\log(2/\eta)^{1/4}, 1\}$. Moreover, assume the bound to hold uniformly with respect to λ, meaning that the collection of training sets for which it holds with confidence $1-\eta$ does not depend on λ.

We note that the above assumption essentially stands on the observation that for functions in the RKH space, the following equality holds $\|f\|_{\rho} = \left\|T^{1/2}f\right\|_{\mathcal{H}}$ (see the proof above), so that estimates in the two norms are highly related. Assumption 2 and Proposition 1 immediately yields the following result.

Corollary 1. *If Assumption 2 holds then with probability at least $1-\eta$*

$$\left\|f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}\right\|_{\rho_{\mathbf{z}}} \leq \alpha(\eta)\hat{C}\sqrt{\lambda}\left(\frac{1}{\omega(\lambda)\sqrt{n}} + \mathcal{A}(\lambda)\right),$$

with $\hat{C} = 1 + \alpha(\eta)C_{\kappa}$.

Proof. From Proposition 1

$$\left\|f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}\right\|_{\rho_{\mathbf{z}}} \leq \left\|f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}\right\|_{\rho} + \frac{\alpha(\eta)C_{\kappa}}{n^{1/4}}\left\|f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}\right\|_{\mathcal{H}},$$

so that the proof follows plugging (27), (28) in the above inequality and noting that $n^{-1/4} \leq \sqrt{\lambda}$ since, $\lambda \geq n^{-1/2}$. □

We now recall the adaptive parameter choice we are going to consider. Let

$$\lambda_{\rho_{\mathbf{z}}} = \max\{\lambda_i : \left\|f_{\mathbf{z}}^{\lambda_i} - f_{\mathbf{z}}^{\lambda_j}\right\|_{\rho_{\mathbf{z}}} \leq \frac{4\alpha(\eta)\hat{C}\sqrt{\lambda_j}}{\sqrt{n}\omega(\lambda_j)}, j = 0, 1, \ldots, i\},$$

$$\lambda_{\mathcal{H}} = \max\{\lambda_i : \left\|f_{\mathbf{z}}^{\lambda_i} - f_{\mathbf{z}}^{\lambda_j}\right\|_{\mathcal{H}} \leq \frac{4\alpha(\eta)}{\sqrt{n}\omega(\lambda_j)}, j = 0, 1, \ldots, i\},$$

and

$$\hat{\lambda} = \min\{\lambda_{\rho_{\mathbf{z}}}, \lambda_{\mathcal{H}}\}. \tag{29}$$

Theorem 3. *Assume that Assumption 2 holds. Let $\lambda_{start} \leq 1/(c\sqrt{n})$. If*

$$\omega(\lambda_{i+1}) \leq q\omega(\lambda_i), \tag{30}$$

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

$$\left\| f_{\mathbf{z}}^{\hat{\lambda}} - f_{\mathcal{H}} \right\|_{\rho} \leq qC\alpha(\eta)\lambda_o(n)\mathcal{A}(\lambda_o(n)),$$

where the value of C can be explicitly given.

Proof. We previously note a few useful facts. Let $\Theta(\lambda) = \omega(\lambda)\mathcal{A}(\lambda)$. First, from Assumption 2- item 3, if we take $\lambda = \lambda_o(n)$ we have

$$\Theta(\lambda_o(n)) \leq c\lambda_o(n) \Rightarrow \frac{1}{\sqrt{n}} \leq c\lambda_o(n) \Rightarrow \frac{1}{n^{1/4}} \leq c\sqrt{\lambda_o(n)}. \tag{31}$$

Second, noting that (30) implies $\omega(\lambda_{i+1})/\sqrt{\lambda_{i+1}} \leq q\omega(\lambda_i)/\sqrt{\lambda_i}$ and recalling the reasoning to get (21), we have

$$\frac{\sqrt{\lambda_*}}{\omega(\lambda_*)} \leq \frac{q\sqrt{\lambda_o(n)}}{\omega(\lambda_o(n))}.$$

This immediately yields

$$\frac{1}{\omega(\lambda_{\rho_{\mathbf{z}}})} \leq \frac{q}{\omega(\lambda_o(n))}, \tag{32}$$

since $\lambda_{\rho_{\mathbf{z}}} \geq \lambda_*$, and

$$\frac{\sqrt{\lambda_{\mathcal{H}}}}{\omega(\lambda_{\mathcal{H}})} \leq \frac{q\sqrt{\lambda_o(n)}}{\omega(\lambda_o(n))}, \tag{33}$$

since $\lambda_{\mathcal{H}} \geq \lambda_*$ and $\sqrt{\lambda}\omega\lambda$ is a decreasing function.

We now consider the two cases: $\lambda_{\rho_{\mathbf{z}}} < \lambda_{\mathcal{H}}$ and $\lambda_{\rho_{\mathbf{z}}} > \lambda_{\mathcal{H}}$.

Case 1. First, consider the case $\hat{\lambda} = \lambda_{\rho_{\mathbf{z}}} < \lambda_{\mathcal{H}}$. From Proposition 1 we have

$$\begin{aligned} \left\| f_{\mathbf{z}}^{\hat{\lambda}} - f_{\mathcal{H}} \right\|_{\rho} &\leq \left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathcal{H}} \right\|_{\rho_{\mathbf{z}}} + \frac{\alpha(\eta)C_\kappa}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathcal{H}} \right\|_{\mathcal{H}} \\ &\leq \left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathcal{H}} \right\|_{\rho_{\mathbf{z}}} + \frac{\alpha(\eta)C_\kappa}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} \right\|_{\mathcal{H}} + \\ &\quad \frac{\alpha(\eta)C_\kappa}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathcal{H}} \right\|_{\mathcal{H}}. \end{aligned} \tag{34}$$

We consider the various terms separately. Applying Theorem 1 and Corollary 1 we get

$$\left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathcal{H}} \right\|_{\rho_{\mathbf{z}}} \leq 6q\alpha(\eta)\hat{C}\sqrt{\lambda_o(n)}\mathcal{A}(\lambda_o(n)). \tag{35}$$

Applying again Theorem 1 and with aid of (31) we also have

$$\frac{\alpha(\eta)C_\kappa}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathcal{H}} \right\|_{\mathcal{H}} \leq 6q\alpha(\eta)^2 cC_\kappa\sqrt{\lambda_o(n)}\mathcal{A}(\lambda_o(n)). \tag{36}$$

Recalling the definition of $\lambda_{\mathcal{H}}$ we also have

$$\left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} \right\|_{\mathcal{H}} \leq \frac{4\alpha(\eta)}{\sqrt{n}\omega(\lambda_{\rho_{\mathbf{z}}})}. \tag{37}$$

We can now use (31), (32) and the definition of $\lambda_o(n)$ to get

$$\frac{\alpha(\eta)C_\kappa}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} \right\|_{\mathcal{H}} \leq 4q\alpha(\eta)^2 c C_\kappa \sqrt{\lambda_o(n)} \mathcal{A}(\lambda_o(n)). \tag{38}$$

If we now plug (35), (36), (38) in (34) we get

$$\left\| f_{\mathbf{z}}^{\hat{\lambda}} - f_{\mathcal{H}} \right\|_\rho \leq q\alpha(\eta) C \sqrt{\lambda_o(n)} \mathcal{A}(\lambda_o(n)),$$

with $C = 6\hat{C} + 10\alpha(\eta)cC_\kappa$.

Case 2. Consider the case $\hat{\lambda} = \lambda_{\mathcal{H}} < \lambda_{\rho_{\mathbf{z}}}$. From Proposition 1 we have

$$\begin{aligned} \left\| f_{\mathbf{z}}^{\hat{\lambda}} - f_{\mathcal{H}} \right\|_\rho &\leq \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathcal{H}} \right\|_{\rho_{\mathbf{z}}} + \frac{\alpha(\eta)C_\kappa}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathcal{H}} \right\|_{\mathcal{H}} \\ &\leq \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} \right\|_{\rho_{\mathbf{z}}} + \left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathcal{H}} \right\|_{\rho_{\mathbf{z}}} + \\ &\quad \frac{\alpha(\eta)C_\kappa}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathcal{H}} \right\|_{\mathcal{H}} \end{aligned} \tag{39}$$

Applying Theorem 1 and using (31) we immediately get

$$\frac{\alpha(\eta)C_\kappa}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathcal{H}} \right\|_{\mathcal{H}} \leq 6qc\alpha(\eta)^2 C_\kappa \sqrt{\lambda_o(n)} \mathcal{A}(\lambda_o(n)). \tag{40}$$

Another straightforward application of Theorem 1 and Corollary 1 gives

$$\left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathcal{H}} \right\|_{\rho_{\mathbf{z}}} \leq 6q\alpha(\eta)\hat{C} \sqrt{\lambda_o(n)} \mathcal{A}(\lambda_o(n)). \tag{41}$$

Finally we have from the definition of $\lambda_{\rho_{\mathbf{z}}}$

$$\left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} \right\|_{\rho_{\mathbf{z}}} \leq \frac{4\alpha(\eta)\hat{C}\sqrt{\lambda_{\mathcal{H}}}}{\sqrt{n}\omega(\lambda_{\mathcal{H}})}, \tag{42}$$

so that using (33), (31) and the definition of $\lambda_o(n)$ we can write

$$\left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} \right\|_{\rho_{\mathbf{z}}} \leq 4\alpha(\eta)q\hat{C} \sqrt{\lambda_o(n)} \mathcal{A}(\lambda_o(n)). \tag{43}$$

The proof is finished plugging (40), (41) and (43) in (39) to get

$$\left\| f_{\mathbf{z}}^{\hat{\lambda}} - f_{\mathcal{H}} \right\|_\rho \leq \alpha(\eta) q C \sqrt{\lambda_o(n)} \mathcal{A}(\lambda_o(n)),$$

where $C = 6\alpha(\eta)C_\kappa + 10\hat{C}$. □

5. Numerical experiments

In this section we consider some numerical experiments discussing how the balancing principle can be approximatively implemented in the presence of very small samples. When the number of samples is very small, as it is often the case in practice, we observed that one cannot completely rely on the theoretical constructions since the bound are conservative and tend to select a large parameter which will oversmooth the estimator. For our numerical experiments, besides the standard regularized least square algorithm, we consider also the more complex situation when the kernel is not fixed in advance but is found within the regularization procedure. We first give a brief summary of this latter approach. Indeed, once a regularized kernel based learning method is applied, two questions should be answered. One of them is how to choose a regularization parameter. The balancing

principle discussed in previous sections provides an answer to this question. Another question is how to choose the kernel, since in several practically important applications a kernel is not a priori given. This question is much less studied. It has been discussed recently in [29], where it has been suggested to select a kernel $K = K(\lambda)$ from some set $\mathbb{K}$ such that

$$K(\lambda) = \arg\min\{Q_{\mathbf{z}}(K,\lambda), K \in \mathbb{K}\}, \tag{44}$$

where

$$Q_{\mathbf{z}}(K,\lambda) = \min_{f\in\mathcal{H}_K} (\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\|f\|^2_{\mathcal{H}_K}),$$

and $\mathcal{H}_K$ is the RKH space generated by K. By definition selected kernel $K = K(\lambda)$ is λ-dependent, so that this kernel choice rule is only applicable for a priori given regularization parameter λ.

At the same time, under rather general assumptions [4] the best in the model $f_{\mathcal{H}_K} \in \mathcal{H}_K$ can be approximated by minimizers $f^{\lambda}_{\mathbf{z}} \in \mathcal{H}_K$ of $Q_{\mathbf{z}}(K,\lambda)$ in such a way that Assumption 2 is satisfied. Then in accordance with the Theorem 3 the parameter choice rule $\lambda = \hat{\lambda} = \hat{\lambda}(K)$ given by (29) allows an accuracy which is only by a constant factor worse than optimal one for fixed $K \in \mathbb{K}$.

Let Λ: $\mathbb{R}_+ \to \mathbb{R}_+$ be the function such that its value at point λ is the number $\hat{\lambda} = \hat{\lambda}(K(\lambda))$ calculated in accordance with (29) for estimators based on the kernel $K(\lambda) \in \mathbb{K}$ given by (44). If $\mathring{\lambda}$ is a fix point of Λ, i.e. $\mathring{\lambda} = \hat{\lambda}(K(\mathring{\lambda}))$, then $K(\mathring{\lambda})$ can be seen as the kernel of optimal choice in the sense of [29], since it satisfies the criterion $Q_{\mathbf{z}}(K,\lambda) \to \min$ for the regularization parameter $\lambda = \mathring{\lambda}$, which is order-optimal for this kernel.

The existence of this fixed point $\lambda = \mathring{\lambda}$ depends on the set $\mathbb{K}$, and deserves consideration in the future. In computational experiment below we find such fix point numerically for an academic example from [29]. At this point it is worth to note that parameter choice rule (29) can be capacity independent in a sense that it does not require a knowledge of spectral properties of underlying kernel K. This feature of the rule (29) makes its combination with the rule (44) numerically feasible.

To simplify a numerical realization of the rule (29) and especially in the presence of very small samples, one can approximate the values $\lambda_{\rho_{\mathbf{z}}}$, $\lambda_{\mathcal{H}}$ using well-known quasi-optimality criterion [37]. As it was observed in [30] this criterion can be seen as a heuristic counterpart of the parameter choice rule $\lambda = \bar{\lambda}$ theoretically justified by Theorem 2. It also operates with norms $\sigma(j) = \|f^{\lambda_j}_{\mathbf{z}} - f^{\lambda_{j-1}}_{\mathbf{z}}\|$, $\lambda_j = \lambda_{start}\cdot\mu^j$, and selects $\lambda^{q-0} = \lambda_l$ such that for any $j = 1,2,...,N$. $\sigma(j) \geq \sigma(l)$, i.e.

$$l = \arg\min\{\sigma(j), j = 1,2,...,N\}.$$

In our experiments we approximate $\lambda_{\rho_{\mathbf{z}}}$ and $\lambda_{\mathcal{H}}$ by

$$\lambda^{q-0}_{\rho_{\mathbf{z}}} = \lambda_l, \quad l = \arg\min\{\sigma_{\rho_{\mathbf{z}}}(j) = \|f^{\lambda_j}_{\mathbf{z}} - f^{\lambda_{j-1}}_{\mathbf{z}}\|_{\rho_{\mathbf{z}}}, \quad j = 1,2,...,N\},$$

and

$$\lambda^{q-0}_{\mathcal{H}} = \lambda_m, \quad m = \arg\min\{\sigma_{\mathcal{H}}(j) = \|f^{\lambda_j}_{\mathbf{z}} - f^{\lambda_{j-1}}_{\mathbf{z}}\|_{\mathcal{H}}, \quad j = 1,2,...,N\},$$

respectively. Then in accordance with (29) we choose a regularization parameter

$$\hat{\lambda} = \min\{\lambda^{q-0}_{\rho_{\mathbf{z}}}, \lambda^{q-0}_{\mathcal{H}}\}. \tag{45}$$

As in [29], we consider a target function

$$f_\rho(x) = \frac{1}{10}(x + 2(e^{-8(\frac{4}{3}\pi - x)^2} - e^{-8(\frac{\pi}{2} - x)^2} - e^{-8(\frac{3}{2}\pi - x)^2})). \quad x \in [0, 2\pi], \tag{46}$$

and a training set $\mathbf{z} = \mathbf{z}_n = \{(x_i, y_i)\}_{i=1}^n$, where $x_i = \frac{2\pi(i-1)}{n-1}$, $y_i = f_\rho(x_i) + \zeta_i$, and ζ_i are random variables uniformly sampled in the interval [-0.02, 0.02].

In our first experiment we test approximate version (45) of the rule (29) using a priori information that the target function (46) belongs to RKH space $\mathcal{H} = \mathcal{H}_K$ generated by the kernel $K(x,t) = K_\rho(x,t) = xt + e^{-8(t-x)^2}$, $t, x \in [0, 2\pi]$.

Figure 2 and 3 display the values $\sigma_{\rho_\mathbf{z}}(j)$, $\sigma_\mathcal{H}(j)$ calculated for regularized least squares estimators $f_\mathbf{z}^{\lambda_j}$, which are constructed using the kernel K_ρ for training sets $\mathbf{z} = \mathbf{z}_{21}$ and $\mathbf{z} = \mathbf{z}_{51}$ respectively. Here and in the next experiment

$$\lambda_j \in \{\lambda_{start} \cdot \mu^j, j = 1, 2, ..., 20\}, \quad \lambda_{start} = 10^{-6}, \mu = 1.5.$$

It is instructive to see that the sequences $\sigma_{\rho_\mathbf{z}}(j)$, $\sigma_\mathcal{H}(j)$, $j = 1, 2, ..., 20$, exhibit different behavior for training sets $\mathbf{z}_{21}$ and $\mathbf{z}_{51}$. At the same time, they attain the minimal values at the same j. Therefore, in accordance with the rule (45) we take $\hat\lambda = \lambda_{\rho_\mathbf{z}}^{q-0} = \lambda_\mathcal{H}^{q-0} = 1.5 \cdot 10^{-6}$ in case of $\mathbf{z} = \mathbf{z}_{21}$, while for $\mathbf{z} = \mathbf{z}_{51}$ $\hat\lambda = \lambda_{\rho_\mathbf{z}}^{q-0} = \lambda_\mathcal{H}^{q-0} = 0.0033$.

Figure 4 and 5 show that for chosen values of parameters the estimator $f_\mathbf{z}^{\hat\lambda}$ provides an accurate reconstruction of the target function.

In our second experiment we do not use a priori knowledge of the space $\mathcal{H}_K$, $K = K_\rho$, containing the target function (46). Instead, we choose a kernel K adaptively from the set.

$$\mathbb{K} = \{K(x,t) = (xt)^\beta + e^{-\gamma(x-t)^2}, \beta \in \{0.5, 1, ..., 4\}, \gamma \in \{1, 2, ..., 10\}\}.$$

trying to find a fix point of the function Λ: $\lambda \to \hat\lambda(K(\lambda))$, where $\hat\lambda(K(\lambda))$ is the number (45) calculated for the kernel $K(\lambda)$, which minimizes $Q_\mathbf{z}(K, \lambda)$, $\mathbf{z} = \mathbf{z}_{21}$, over the set $\mathbb{K}$.

In the experiment we take $\lambda^{(s)} \in \{\lambda_j\}_{j=1}^{20}$ and find the minimizer $K(\lambda^{(s)}) \in \mathbb{K}$ by the simple full-search over the finite set $\mathbb{K}$. Then next value $\lambda^{(s+1)} \in \{\lambda_j\}_{j=1}^{20}$ is defined as the number (45) calculated for estimators $f_\mathbf{z}^{\lambda_j}$ based on the kernel $K(\lambda^{(s)})$. This iteration procedure terminates when $|\lambda^{(s+1)} - \lambda^{(s)}| \le 10^{-4}$. It gives us required approximate fix point $\mathring\lambda = \lambda_{18} \approx 0.0014$ and corresponding kernel $K(\mathring\lambda) = K(\mathring\lambda; x, t) = xt + e^{-10(x-t)^2}$, which is a good approximation for the ideal kernel $K_\rho(x,t)$. The estimator $f_\mathbf{z}^{\mathring\lambda}$ based on the kernel $K(\mathring\lambda)$ provides a good reconstruction of the target function (46), as it can be seen in the Figure 6.

Presented numerical experiments demonstrate a reliability of the balancing principle, and show that it can be used also in learning the kernel function via regularization.

Acknowledgments

This research was started when S. Pereverzyev visited DISI, University of Genova. Many thanks for the hospitality and excellent working conditions. The work of S. Pereverzyev is partially supported by EU-project "DIAdvisor" performed within 7-th Framework Programme of EC. Ernesto De Vito and Lorenzo Rosasco have

been partially supported by the FIRB project RBIN04PARL and by the the EU Integrated Project Health-e-Child IST-2004-027749. The numerical simulations presented in the Section 5 were carried out in MATLAB, and they are reproduced here with kind permission by Huajun Wang, RICAM, Linz.

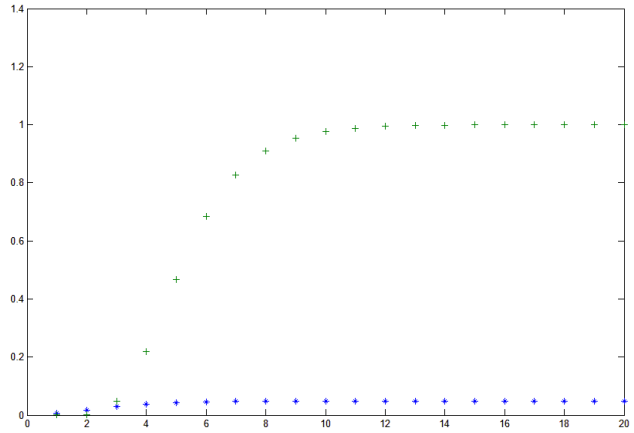

Figure 2: The values of $\sigma_{\rho_{\mathbf{z}}}(j)$ (blue dots) and $\sigma_{\mathcal{H}}(j)$ (green crosses) for $\mathbf{z} = \mathbf{z}_{21}$.

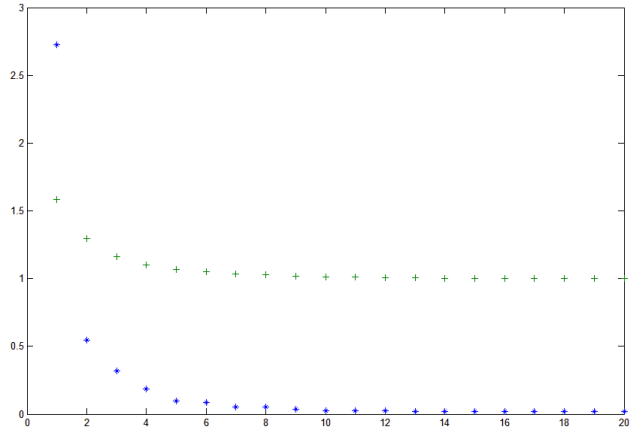

Figure 3: The values of $\sigma_{\rho_{\mathbf{z}}}(j)$ (blue dots) and $\sigma_{\mathcal{H}}(j)$ (green crosses) for $\mathbf{z} = \mathbf{z}_{51}$

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[2] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[3] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pages 286–297, 2000.

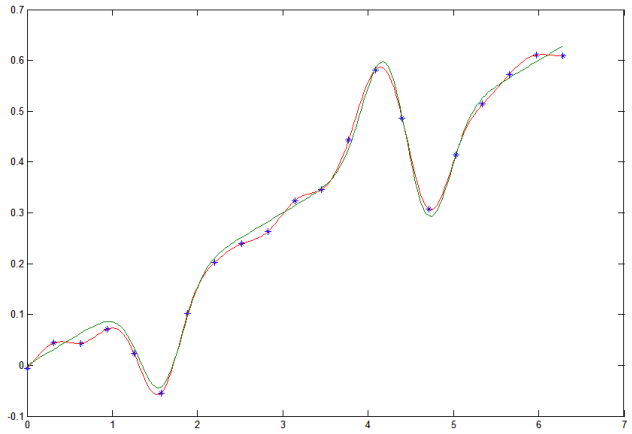

Figure 4: The estimator $f_{\mathbf{z}}^{\hat{\lambda}}$ (red line) and the target function f_ρ (green line) for $\hat{\lambda} = 1.5 \cdot 10^{-6}$ and training set $\mathbf{z} = \mathbf{z}_{21}$ (blue dots).

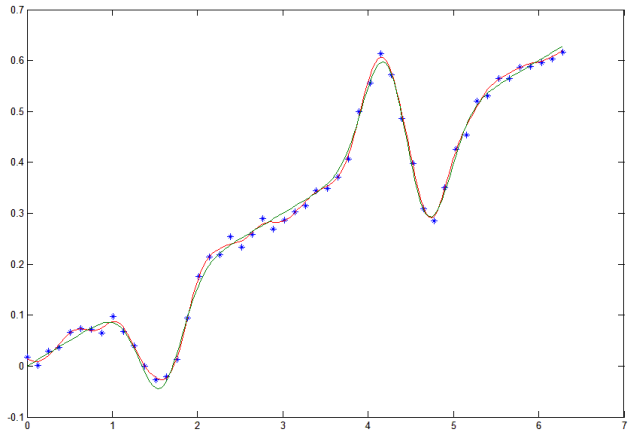

Figure 5: The estimator $f_{\mathbf{z}}^{\hat{\lambda}}$ (red line) and the target function f_ρ (green line) for $\hat{\lambda} = 0.0033$ and training set $\mathbf{z} = \mathbf{z}_{51}$ (blue dots).

[4] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23(1):52–72, 2007.

[5] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375 (electronic), 2005.

[6] P. Bühlmann and B. Yu. Boosting with the l_2-loss: Regression and classification. *Journal of American Statistical Association*, 98:324–340, 2002.

[7] A. Caponnetto. Optimal rates for regularization operators in learning theory. Technical report, CBCL Paper #264/ CSAIL-TR #2006-062, M.I.T, 2006. available at http://cbcl.mit.edu/projects/cbcl/publications/ps/MIT-CSAIL-TR-2006-062.pdf.

[8] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.

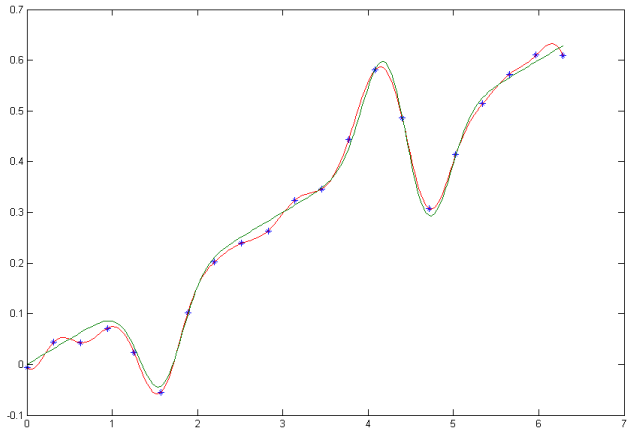

Figure 6: The target function f_ρ (green line) and its estimator $f_{\mathbf{z}}^{\mathring{\lambda}}$ (red line) based on the adaptively chosen kernel $K(\mathring{\lambda}; x, t) = xt + e^{-10(x-t)^2}$, $\mathring{\lambda} = 0.0014$, and training set $\mathbf{z} = \mathbf{z}_{21}$ (blue dots).

[9] A. Caponnetto and Y. Yao. Adaptation for regularization operators in learning theory. Technical Report CBCL Paper 265, CSAIL-TR 2006-063, Massachusetts Institute of Technology, Cambridge, MA, 2006. submitted for publication.

[10] Andreas Christmann and Ingo Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.

[11] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

[12] C. De Mol, E. De Vito, and L. Rosasco. Elastic net regularization in learning theory. Technical Report CBCL paper 273/ CSAILTechnical Report TR-2008-046, Massachusetts Institute of Technology, 2008.

[13] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, May 2005.

[14] E. De Vito, L. Rosasco, and A. Verri. Spectral methods for regularization in learning theory. Technical Report Technical Rerport DISI-TR-05-18., DISI, Universitá degli Studi di Genova, Italy, 2005.

[15] R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov. On mathematical methods of learning. *Foundations of Computational Mathematics*, 6(1):3–58, 2006.

[16] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.

[17] Sandrine Dudoit and Mark J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat. Methodol.*, 2(2):131–154, 2005.

[18] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[19] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Adv. Comp. Math.*, 13:1–50, 2000.

[20] Alexander Goldenshluger and Sergei V. Pereverzev. On adaptive inverse estimation of linear functionals in Hilbert scales. *Bernoulli*, 9(5):783–807, 2003.

[21] M. Györfi, L.and Kohler, A. Krzyzak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer Series in Statistics, New York, 1996, 2002.

[22] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[23] V. Koltchisnkii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2004.

[24] O. Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probab. and Its Appl.*,, 35:454–466, 1990.
[25] Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. *Ann. Statist.*, 32(4):1679–1697, 2004.
[26] Peter Mathé. The Lepskiĭ principle revisited. *Inverse Problems*, 22(3):L11–L15, 2006.
[27] Peter Mathé and Sergei V. Pereverzev. Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(3):789–803, 2003.
[28] Peter Mathé and Sergei V. Pereverzev. Regularization of some linear ill-posed problems with discretized random noisy data. *Math. Comp.*, 75(256):1913–1929 (electronic), 2006.
[29] Charles A. Micchelli and Massimiliano Pontil. Learning the kernel function via regularization. *J. Mach. Learn. Res.*, 6:1099–1125 (electronic), 2005.
[30] Sergei Pereverzev and Eberhard Schock. On the adaptive selection of the parameter in regularization of ill-posed problems. *SIAM J. Numer. Anal.*, 43(5):2060–2076 (electronic), 2005.
[31] I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for probabilities of large deviations. *Theory Probab. Appl.*, 30(1):143–148, 1985.
[32] Tomaso Poggio and Federico Girosi. A theory of networks for learning. *Science*, (247):978–982, 1990.
[33] L. Rosasco. *Regularization Approaches in Learning Thoery*. PhD Thesis, University of Genova, 2006.
[34] R. Rosipal, L.J. Trejo, and A. Cichocki. Kernel principal component regression with em approach to nonlinear principal components extraction. Technical report, University of Paisley, 2000.
[35] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
[36] S. Smale and D.X. Zhou. Learning theory estimates via integral operators and their approximations. *submitted to Constructive Approximation*, 2005. retrievable at http://www.tti-c.org/smale.html.
[37] A.N. Tikhonov and V.B. Glasko. Use of the regularization method in non-linear problems. *Zh.vychisl.Mat.mat.Fiz.*, 5:463–473, 1965.
[38] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.
[39] Aad W. van der Vaart, Sandrine Dudoit, and Mark J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statist. Decisions*, 24(3), 2006.
[40] V. N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.
[41] Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Learning rates of least-square regularized regression. *Found. Comput. Math.*, 6(2):171–192, 2006.
[42] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constr. Approx.*, 26(2):289–315, 2007.
[43] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *JRSSB*, 67(2):301–320, 2005.

Ernesto De Vito, D.S.A., Università di Genova and INFN, Sezione di Genova
E-mail address: devito@dima.unige.it

Johann Radon Institute for Computational and Applied Mathematics,Austrian Academy of Sciences, Altenbergerstrasse 69, A-4040 Linz, Austria
E-mail address: sergei.pereverzyev@oeaw.ac.at

Lorenzo Rosasco, Center for Biological and Computational Learning at the Massachusetts Institute of Technology, & DISI, Università di Genova, Italy
E-mail address: lrosasco@mit.edu