Found Comput Math (2010) 10: 455–479
DOI 10.1007/s10208-010-9064-2

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Adaptive Kernel Methods Using the Balancing Principle

E. De Vito · S. Pereverzyev · L. Rosasco

Received: 16 October 2008 / Revised: 31 August 2009 / Accepted: 28 January 2010 /
Published online: 19 March 2010

Abstract The regularization parameter choice is a fundamental problem in Learning Theory since the performance of most supervised 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 needed to choose the regularization parameter in order to obtain good learning rates. In this paper we present a parameter choice strategy, called 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

Communicated by Felipe Cucker.

E. De Vito
DSA, Università di Genova and INFN, Genova, Italy
e-mail: devito@dima.unige.it

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

L. Rosasco (✉)
Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: lrosasco@mit.edu

L. Rosasco
DISI, Università di Genova, Genova, Italy

can immediately derive adaptive parameter choices for various kernel methods recently studied. Numerical experiments with the proposed parameter choice rules are also presented.

Keywords Learning Theory · Model Selection · Adaptive Regularization · Inverse Problems

Mathematics Subject Classification (2000) Primary 68T05 · 68Q32

1 Introduction

Most supervised learning algorithms depend on some tuning parameter, whose correct choice is crucial to ensure good performances of the solution. Examples are the regularization parameter in regularized least-squares regression [20] or the complexity of the hypothesis space in empirical risk minimization [41]. The error incurred by a learning algorithm is usually the sum of two terms, sample and approximation errors, having opposite behavior with respect to the tuning parameter [11]—see Fig. 1. In this context a natural parameter choice is obtained via a trade-off between sample and approximation error. This choice is shown to often provide optimal convergence rates in a mini-max setting [4, 8, 11, 23, 35] and we refer to it as the *best parameter choice*.

The above strategy raises conceptual and practical issues since estimates of the approximation error depend on a priori knowledge on the problem which is usually not available. The so-called *no free lunch theorem* shows that no data-independent parameter choice can achieve the best convergence rate [23]. To overcome this problem, a data-driven choice is needed, ensuring the error rate of the solution achieves the best possible rate. In the statistical literature this problem is known as the problem of adaptive model selection [16, 23]. In regression models with fixed design, classical model selection schemes include the Akaike criterion and BIC among others (see [24] 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 [2, 5, 17, 23] for general references and also [3, 27]), on data splitting, e.g., hold-out and cross-validation (see [17] and more recently [9, 18, 40]) and on aggregation [39]. In particular the application of aggregation techniques to select regularization parameters is recently discussed in [21]. In this paper we are interested in regularization parameter choices that do not require any data splitting.

Based on the relation between learning theory and the theory of regularization in inverse problems—see [14, 20, 31, 34, 41] and references therein—in this paper we study a data-driven method for a regularization parameter choice, namely *the balancing principle*. This method is a development of an approach proposed by [26] in the context of Gaussian regression, and has been studied in the context of inverse problems in [22] and eventually developed in a series of papers (see [28] and references therein). Related approaches have been considered in statistical learning for aggregation of classifiers [39] and for empirical risk minimization algorithms [25]. Here we extend the approach proposed in [22] which is very natural when considering regularized kernel methods.

The approaches to a posteriori parameter choice in inverse problems cannot be used directly in the context of learning, since they are based on estimates of the stability of regularization methods as measured in the space where the element of interest (regression or target function) should be recovered. In the context of learning theory, typically, such estimates are measured with respect to the expected risk which depends on the unknown probability measure. In this paper we discuss how the inverse problems results can be adapted to the learning setting. The method we introduce is simple, requires no data splitting, and adaptively achieves the best possible error rate. The proposed method allows to easily derive adaptive parameter choices achieving optimal rates for several kernel methods [4, 7, 8, 35, 43] and we believe it might serve as a general way to obtain adaptive regularization schemes in kernel spaces.

The paper is organized as follows. In Sect. 2 we give the necessary background on supervised learning and discuss in some detail the problem of adaptive regularization parameter choice. In Sect. 3 we present and discuss our main results, while the proofs are postponed to Sect. 4. We conclude in Sect. 5 with some numerical experiments.

2 Regularized Learning and Adaptive Parameter Choice

In this section, after recalling the basic concepts in supervised learning, we discuss the problem of adaptive regularization parameter choice motivating the study in this paper.

2.1 Some Background on Supervised Learning

We consider the problem of supervised learning [11, 41]. Given a *training set*, $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = (x_1, y_1), \ldots, (x_n, y_n)$, the goal is to find an input–output relation $f : X \to Y$. More precisely, $x \in X \subset \mathbb{R}^d$, $y \in Y \subseteq \mathbb{R}$, and the data are sampled identically and independently from an unknown probability measure ρ on $X \times Y$. For a chosen *loss function* $\ell : X \times \mathbb{R} \to \mathbb{R}^+$, the error incurred by a function is measured by the *expected risk*

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

and in this paper we are primarily interested in the square loss $\ell(y, f(x)) = (y - f(x))^2$. The search for a possible estimate is typically restricted to a *hypotheses space* $\mathcal{H}$, e.g., splines [42] or reproducing kernel Hilbert (RKH) spaces [1, 11]. In this case the ideal solution is the (so-called) *best in the model*[1] $f_{\mathcal{H}}$ such that

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

This solution cannot be computed in practice since ρ is unknown and a learning algorithm can be seen as a map $\mathbf{z} \to f_{\mathbf{z}} \in \mathcal{H}$, that given a training set provides us with

[1] 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.

an *estimator* $f_{\mathbf{z}}$ of $f_{\mathcal{H}}$. To assess the quality of an estimator we need to fix: an error measure to quantify how well $f_{\mathbf{z}}$ approximates $f_{\mathcal{H}}$; some probabilistic tools, since $f_{\mathbf{z}}$ is a random variable.

We will measure errors 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 [38]. This latter case is of interest for estimators in RKHS since estimates in various other norms, e.g., uniform norm or Sobolev norms, can be easily obtained—see [35]. Also, it is of interest in sparse learning—see [13] and references therein—where one is interested in estimating the coefficients obtained expanding $f_{\mathcal{H}}$ on a given dictionary.

The error bounds we consider are standard in statistical learning, and correspond to probabilistic inequalities of the form

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

where $0 < \eta \leq 1$. The above inequality is called an excess risk bound. Equivalently, we have $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \leq \varepsilon(n,\eta)$, where the last inequality holds with probability at least $1-\eta$. Similarly, when the error is measured by a fixed norm, one can consider

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

As we mentioned before, it is well known that to obtain finite sample bounds prior assumptions on the problem at hand are required [16, 17, 23]. The impact of this fact on the design of a fully data-driven algorithm is at the basis of the study in this paper. In the next section, we discuss this point of view in detail.

2.2 Regularization Parameter Choice

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-step procedure. The first step defines a family of solutions depending on a real *regularization* parameter $\mathbf{z} \to f_{\mathbf{z}}^{\lambda}$, $\lambda > 0$, whereas the second step determines how to choose the regularization parameter λ. The final estimator is obtained only when both steps are defined. Among other algorithms, regularization networks [20] and support vector machines [41] can be cast in this framework.

One fundamental approach to model selection, in learning theory [5, 12, 36, 41] as well as in non-parametric regression [23], is based on deriving excess risk bounds for any λ, and choose the value optimizing the bound. More precisely, excess risk bounds are usually given by the sum of two competing terms, i.e.,

$$\mathcal{E}\big(f_{\mathbf{z}}^{\lambda}\big) - \mathcal{E}(f_{\mathcal{H}}) \leq \mathcal{S}(n,\eta,\lambda) + \mathcal{A}(\lambda). \tag{3}$$

The term $\mathcal{S}(n,\eta,\lambda)$ is the so-called sample error and quantifies the error due to random sampling. The term $\mathcal{A}(\lambda)$ is the approximation error; it does not depend on the data, but requires prior knowledge on the unknown probability distribution. The typical behavior of the two terms (for fixed n) is depicted in Fig. 1.

The best possible regularization parameter choice is found by solving a sample-approximation (or bias-variance) trade-off, that is from the balancing of these two

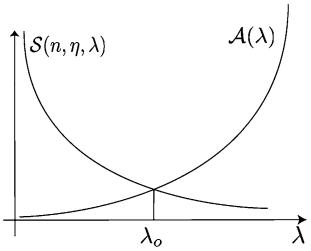

Fig. 1 The figure represents the behavior of sample and approximation errors, respectively $\mathcal{S}(n,\eta,\lambda)$ and $\mathcal{A}(\lambda)$, as functions of λ, for fixed n,η

terms. In this paper, rather than the value optimizing the bound, we consider the value of $\lambda_o(n)$ making the contribution of the two terms equal[2] (the crossing point in Fig. 1). One can see that the corresponding error estimate is, with probability at least $1-\eta$

$$\mathcal{E}\big(f_{\mathbf{z}}^{\lambda_o(n)}\big) - \mathcal{E}(f_{\mathcal{H}}) \leq 2\mathcal{S}(n,\eta,\lambda_o) = 2\mathcal{A}(\lambda_o). \tag{4}$$

Before developing our reasoning further we give an example.

Example 1 (Regularized Least Squares) Consider the regularized least-squares estimator $f_{\mathbf{z}}^{\lambda}$ solving

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

where $\mathcal{H}$ is a RKH space [1] with bounded kernel K and $\|\cdot\|_{\mathcal{H}}$ the corresponding norm. The regularity assumption is that

$$f_{\mathcal{H}} = L_K^r u, \qquad L_K f(x) = \int_X K(x,s) f(s)\,\mathrm{d}\rho_X(s), \tag{5}$$

for some u such that $\int |u(x)|^2\,\mathrm{d}\rho_X(x) < \infty$, where ρ_X denotes the marginal probability of ρ on X. In this case, one can prove [4, 7, 8, 35] that

$$\mathcal{E}\big(f_{\mathbf{z}}^{\lambda}\big) - \mathcal{E}(f_{\mathcal{H}}) \leq C\log\left(\frac{4}{\eta}\right)\left(\frac{1}{\lambda n} + \lambda^{2r}\right), \quad \frac{1}{2} < r \leq 1, \tag{6}$$

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

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

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

From the above discussion, one can see that the parameter choice $\lambda_o(n)$ depends on the regularity properties of $f_{\mathcal{H}}$ that are usually not known. This observation mo-

[2]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.

tivates the interest in the adaptive parameter choice that we study in this paper. We conclude this section with two remarks.

Remark 1 (Optimality and Minimax Results) In this paper we refer to the value $\lambda_o(n)$ 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 is tight. We do not discuss this problem and we refer to [8, 16, 23, 38] for further discussion.

Remark 2 (Optimality and Order Optimality) In our analysis we can usually compute the value of essentially all the constants appearing, but we do not expect such constants to be optimal. For this reason we often take fairly crude estimates, and in fact we mainly focus on recovering the correct dependence on the number of samples. This is related to the difference between order optimality and optimality in inverse problems [19].

3 Adaptive Regularized Learning

In this section we state the main results in the paper. Our main result in Sect. 3.2 deals with adaptive parameter selection for kernel methods when the error is measured via the excess risk, but we first present some auxiliary results 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 right and clarify the main intuition underlying the balancing principle.

3.1 Adaptive Learning when the Error Measure is Known

We assume both the estimator and the best in the model to be elements of some normed space whose norm we denote with $\|\cdot\|$. Such a norm is assumed to be *known* (note that on the contrary the risk is not).

Again we assume that an error bound of the form

$$\|f_{\mathbf{z}} - f_{\mathcal{H}}\| \leq \mathcal{S}(n, \eta, \lambda) + \mathcal{A}(\lambda)$$

is available and further assume that

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

where $\alpha(\eta) > 1$ and ω, γ are positive functions. This latter assumption is typically satisfied and is made only to simplify the exposition. In the case of the regularized least-squares algorithm (see Example 2 below) $\omega(\lambda) = \lambda$, $\gamma(n) = \sqrt{n}$ and $\alpha(\eta) = \log(4/\eta)$. Since $\alpha(\eta) > 1$, we can rewrite the bound as

$$\left\|f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}\right\| \leq \alpha(\eta)\left(\frac{1}{\omega(\lambda)\gamma(n)} + \mathcal{A}(\lambda)\right), \tag{7}$$

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

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

To define a parameter strategy we first consider a suitable discretization for the possible values of the regularization parameter, that is an ordered sequence $(\lambda_i)_{i\in\mathbb{N}}$ such that the best value $\lambda_o(n)$ falls within the considered grid. The balancing principle estimate for $\lambda_o(n)$ is defined via

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

Such an estimate no longer depends on $\mathcal{A}$ and the reason why we can expect it to be still sufficiently close to $\lambda_o(n)$ is better illustrated by Fig. 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 $1 - \eta$

$$\begin{aligned}\left\| f_{\mathbf{z}}^{\alpha} - f_{\mathbf{z}}^{\beta} \right\| &\leq \left\| f_{\mathbf{z}}^{\alpha} - f_{\mathcal{H}} \right\| + \left\| f_{\mathbf{z}}^{\beta} - f_{\mathcal{H}} \right\| \\ &\leq \alpha(\eta)\left(\mathcal{A}(\alpha) + \frac{1}{\gamma(n)\omega(\alpha)}\right) + \alpha(\eta)\left(\mathcal{A}(\beta) + \frac{1}{\gamma(n)\omega(\beta)}\right) \\ &\leq 4\frac{\alpha(\eta)}{\gamma(n)\omega(\alpha)}. \end{aligned} \tag{8}$$

The intuition is that when such a condition is violated we are close to the intersection point of the two curves, that is to $\lambda_o(n)$. The above discussion is made precise in the following.

Assumption 1 For $\lambda \in (0, 1]$ 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 $\alpha(\eta) > 1$, $\gamma(n) > 0$. 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.

It is easy to check that the last item in the above assumption ensures that $\lambda_o(n) \geq 1/(c\gamma(n))$, so that if we choose the first value λ_{start} in the sequence $(\lambda_i)_{i\in\mathbb{N}}$ so that

$\lambda_{\text{start}} \leq 1/(c\gamma(n))$, the best possible parameter choice will fall within the parameter range we consider. The last condition in the assumption requires the bound to be uniform with respect to λ, and is needed since the parameter choice we consider is data dependent. This assumption is satisfied in all the examples of algorithms we consider in the paper. In other cases, such as empirical risk minimization, it can be enforced considering a union bound on the different regularization parameter values—see for example [17], Chap. 18.

The following theorem shows that the value λ_+, given by the balancing principle, provides the same error estimate of $\lambda_{\text{o}}(n)$ up to a constant factor.

Theorem 1 *If Assumption 1 holds and moreover, and we consider a sequence of regularization parameter values such that* $\lambda_{\text{start}} \leq 1/(c\gamma(n))$ *and*

$$\omega(\lambda_{i+1}) \leq q\omega(\lambda_i), \quad q > 1, \tag{9}$$

then with probability at least $1 - \eta$

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

The above theorem shows that the balancing principle can adaptively achieve the best possible learning rate. In its basic formulation the balancing principle requires an extensive comparison of solutions at different values λ_i. In fact, the procedure can be simplified, at the price of slightly worsening the constant in the bound. In fact, we can take a geometric sequence

$$\lambda_i = \lambda_{\text{start}}\mu^i, \quad \text{with } \mu > 1,\ \lambda_{\text{start}} \leq \frac{1}{c\gamma(n)} \tag{10}$$

and introduce the choice

$$\bar{\lambda} = \max\left\{\lambda_i : \left\| f_{\mathbf{z}}^{\lambda_j} - f_{\mathbf{z}}^{\lambda_{j-1}} \right\| \leq \frac{4\alpha(\eta)}{\gamma(n)\omega(\lambda_{j-1})}, j = 1, \ldots, i-1\right\}, \tag{11}$$

requiring only comparison of solutions for adjacent parameter values. The next theorem studies the error estimate obtained with this choice.

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

$$\frac{\omega(2\lambda)}{b} \leq \omega(\lambda) \leq \frac{\omega(2\lambda)}{a}, \tag{12}$$

then, taking a sequence of regularization parameter values as in (10), we have with probability at least $1 - \eta$

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

where C might depend on a, b, μ.

We discuss some cases where the above results apply. The letter C is used to indicate constants independent of λ and n. We first go back to the RLS algorithm and consider error estimates in the RKHS norm.

Example 2 (Regularized Least Squares) Error estimates for the RLS algorithm are known both for the expected risk (see Example 1, (6)) and the RKHS norm [4, 8, 35]. In this latter case with probability $1-\eta$

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

(under the same assumption of Example 1). It is straightforward to check that the above estimate satisfies the conditions needed to apply the balancing principle and achieve the best rate 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, 44] and kernel principal component regression [24, 33]. All such algorithms can be written as

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

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 [4, 7, 15] for details).

The prior assumption (5) 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 (5) restricted to $\mathcal{H}$. The following bound is proved in [4], with probability at least $1-\eta$

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

for any[3] $\lambda \geq n^{-1/2}$.

Example 4 (Elastic-Net Regularization) The elastic-net algorithm proposed in [45], is studied in [13] in the context of learning with an infinite dimensional overcomplete dictionary of features $(\psi_\gamma)_{\gamma\in\Gamma}$. 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 an estimator of the form $\sum_{\gamma\in\Gamma} \beta_\gamma \psi_\gamma$. The idea is that the function of interest is sparse, meaning that many of the coefficients in the previous expansion are zero. In this case a sparse and stable estimator β_n^{λ} is found by minimizing

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

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

where $\epsilon, \lambda > 0$. In this setting one is often interested in error estimates on the coefficients since they give information on which features are most important. 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 [13] that with probability at least $1 - \eta$

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

where $\|\cdot\|_2$ is the norm in $\ell_2(\Gamma)$. Clearly, the above bound satisfies the assumption needed to apply the balancing principle.

Example 5 (Tikhonov Regularization with Convex Loss) The RLS algorithm can be generalized to

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

where $\ell : Y \times \mathbb{R} \to [0, \infty]$ is a loss function which is convex in its second entry. If we denote by f^λ the minimizer of

$$\mathcal{E}(f) + \lambda \|f\|^2_{\mathcal{H}}$$

then, recall that if the loss function is convex, it is also locally Lipschitz continuous so that

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

where the Lipschitz constant L_λ might depend on λ. The following bound is proved in [32] (see also [10] for a more general setting), when the outputs are bounded and the loss is bounded at 0; with probability at least $1 - \eta$

$$\mathcal{E}\big(f^\lambda_{\mathbf{z}}\big) - \mathcal{E}(f_{\mathcal{H}}) \leq C \log\left(\frac{2}{\eta}\right)\left(\frac{L_\lambda}{\lambda\sqrt{n}} + \phi(\lambda)\right),$$

where $\phi(\lambda) = \min_{f \in \mathcal{H}}\{\mathcal{E}(f) + \lambda \|f\|^2_{\mathcal{H}} - \mathcal{E}(f_{\mathcal{H}})\}$. For a large number of loss functions (see [32]) the constant L_λ can be explicitly computed and $\omega(\lambda) = L_\lambda/\lambda$, ϕ satisfy the assumptions required to apply the balancing principle.

3.2 Adaptive Learning for the Expected Risks

Our further goal is 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 $\mathcal{E}(f^{\lambda_i}_{\mathbf{z}}) - \mathcal{E}(f^{\lambda_{i-1}}_{\mathbf{z}})$ and hence a knowledge of the distribution ρ.

To deal with this situation we make two restrictions: (1) we consider regularization algorithms $f^\lambda_{\mathbf{z}}$ in a hypothesis space $\mathcal{H}$ which is a RKH space, (2) we consider regularization algorithms based on the square loss function. 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

$$\kappa = \sup_{x \in X} \sqrt{K(x, x)}. \tag{13}$$

We denote by ρ_X the marginal probability of ρ on X, and by $\rho(y|x)$ the conditional probability. Since we consider the square loss $(y - f(x))^2$, if $\int y^2 \rho(x, y) < \infty$, the expected risk is a well-defined functional on the space $L^2(X, \rho_X) = \{f : X \to \mathbb{R} \mid \|f\|_\rho^2 = \int_X f(x)^2 \, \mathrm{d}\rho_X(x) < \infty\}$. In this case some facts are well known [11, 23]. The minimizer of $\mathcal{E}(f)$ over $L^2(X, \rho_X)$ is the regression function $f_\rho(x) = \int_Y y \, \mathrm{d}\rho(y|x)$ and for $f \in L^2(X, \rho_X)$ we can write

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

Then, as we mentioned before, the application of the balancing principle is not straightforward since we should evaluate $\|f_{\mathbf{z}}^\beta - f_{\mathbf{z}}^\lambda\|_\rho$. Since both the empirical norm

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

and the RKH space norm are known, we can consider

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

and

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

Our main result shows that the choice

$$\hat{\lambda} = \min\{\lambda_{\rho_{\mathbf{z}}}, \lambda_{\mathcal{H}}\} \tag{14}$$

allows to achieve the best error rate for the expected risk in an adaptive way. To show this we need the following assumption.

Assumption 2 Assume that $\lambda \ge n^{-1/2}$ and that the following bounds hold with probability at least $1 - \eta$:

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

and

$$\left\|f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}\right\|_{\mathcal{H}} \le \alpha(\eta)\left(\frac{1}{\sqrt{n}\omega(\lambda)} + \mathcal{A}(\lambda)\right), \tag{16}$$

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.

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, 35]). At the root of this relation there is essentially the fact that the reproducing kernel Hilbert space can be viewed as the image of $L^2(X, \rho_X)$ under $L_K^{-1/2}$ [11]. Because of this fact, an estimator will have different norms (that is lie in spheres with different radius) in $L^2(X, \rho_X)$ and $\mathcal{H}_K$.

Given Assumption 2 the best parameter choice $\lambda_o(n)$ is the same in both cases but the rates are different, in fact we have

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

for the expected risk and

$$\left\| f_{\mathbf{z}}^{\lambda_o(n)} - f_{\mathcal{H}} \right\|_{\mathcal{H}} \leq \alpha(\eta)\mathcal{A}\big(\lambda_o(n)\big) \tag{18}$$

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

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

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 by doing this we would not to match the best error rate for the expected risk, as can be seen comparing (17) and (18).

The following theorem is our main result and studies the property of the choice (14).

Theorem 3 *Assume that Assumption 2 holds. Consider a sequence of regularization parameter values such that* $\lambda_{\text{start}} \leq 1/(c\sqrt{n})$ *and*

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

If we choose $\hat{\lambda}$ *as in* (14)*, then the following bound holds with probability at least* $1 - \eta$

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

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 regularized least-square algorithm in RKHS, as well as for the class of spectral regularization algorithms studied in [4, 15]. To the best of our knowledge the balancing principle is the first strategy that allows to achieve this result without requiring any data splitting. A hold-out strategy is discussed in [9], where adaptation is proved also when $f_\rho \neq \mathcal{H}$. Recently the use of aggregation to choose the regularization parameter for the regularized least squares has been considered in [21]. The results in that paper apply to polynomial type of approximation rates and require the knowledge of the spectrum asymptotic for the operator L_K, which depends on the unknown marginal distribution.

We proceed illustrating the application of the above result to the regularized least-square algorithm.

Example 6 (Regularized Least Squares) As we previously mentioned, for the regularized least-square algorithm (see Examples 1 and 2) we have with probability at least $1-\eta$

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

but also

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

Applying the above result we have that the parameter choice (14) satisfies with probability at least $1-\eta$

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

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 $1-\eta$

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

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 that shows how to practically compute (14).

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

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

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

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

Similarly one can see that

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

4 Proofs

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

4.1 Results for Known Norm

Recall that if Assumption 1 holds the best parameter choice achieves the error estimate (4) and it can be shown that the last condition in Assumption 1 ensures $\lambda_{\mathrm{o}}(n) \geq 1/(c\gamma(n))$. Note that, if we now restrict our attention to some discrete sequence $(\lambda_i)_i$ with $\lambda_{\mathrm{start}} \leq 1/(c\gamma(n))$, then it is easy to see that the best estimate for $\lambda_{\mathrm{o}}(n)$ is

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

which still depends on $\mathcal{A}$. Given these observations we can give the proof of Theorem 1.

Proof of Theorem 1 Note that all the inequalities in the proof are to be interpreted as holding with probability at least $1 - \eta$. Recall that by (8) for λ, β such that $\lambda \leq \beta \leq \lambda_{\mathrm{o}}(n)$ 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 $\lambda_* \leq \lambda_{\mathrm{o}}(n)$, and we know that, for any $\lambda_j \leq \lambda_* \leq \lambda_{\mathrm{o}}(n)$,

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

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

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

Finally to relate λ_* and $\lambda_{\mathrm{o}}(n)$, we let $\lambda_* = \lambda_\ell$ so that $\lambda_* = \lambda_\ell \le \lambda_{\mathrm{o}}(n) \le \lambda_{\ell+1}$. Since $\omega(\lambda)$ is increasing, we can use (9) to get $\omega(\lambda_{\mathrm{o}}(n)) \le \omega(\lambda_{\ell+1}) \le q\omega(\lambda_\ell) = q\omega(\lambda_*)$. The above reasoning yields

$$\frac{1}{\omega(\lambda_*)} \le \frac{q}{\omega(\lambda_{\mathrm{o}}(n))}, \tag{21}$$

and if we plug the above inequality into (20), the definition of $\lambda_{\mathrm{o}}(n)$ gives

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

so that the theorem is proved. □

The proof of Theorem 2 is similar.

Proof of Theorem 2 The proof follows exactly the one for deterministic inverse problems though the inequalities here are to be interpreted as holding 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 $m \ge \ell$ and we can use the definition of $\bar{\lambda}$ to write

$$\begin{aligned}\left\| f_{\mathbf{z}}^{\bar{\lambda}} - f_{\mathbf{z}}^{\lambda_*} \right\| &\le \sum_{j=\ell+1}^{m} \left\| f_{\mathbf{z}}^{\lambda_j} - f_{\mathbf{z}}^{\lambda_{j-1}} \right\| \\ &\le 4\alpha(\eta)\frac{1}{\gamma(n)} \sum_{j=\ell+1}^{m} \frac{1}{\omega(\lambda_{j-1})} \\ &\le 4\alpha(\eta)\frac{1}{\gamma(n)} \sum_{j=0}^{m-\ell-1} \frac{1}{\omega(\lambda_*\mu^j)}. \end{aligned} \tag{22}$$

Now for any $\mu > 1$, $\alpha > 1$ let $p, s \in \mathbb{N}$ be such that $2^p \le \mu \le 2^{p+1}$ and $2^s \le \alpha \le 2^{s+1}$. Then using (12) we get

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

The last inequality shows that (9) 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

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

The theorem is proved recalling the definition of $\lambda_{\mathrm{o}}(n)$. □

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. The following concentration result will be crucial.

Proposition 1 *Assume that $\mathcal{H}$ is a RKH space with bounded kernel (13). For $f \in \mathcal{H}$ we have with probability at least $1 - \eta$*

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

and $C_\kappa^2 = 2\sqrt{2}\kappa^2$.

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

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

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 $\forall f \in \mathcal{H}$

$$\left| \|f\|_{\rho} - \|f\|_{\rho_{\mathbf{z}}} \right| \leq \sqrt{\|T - T_{\mathbf{x}}\|}\, \|f\|_{\mathcal{H}}. \tag{23}$$

The quantity $\|T - T_{\mathbf{x}}\|$ have been studied in [4, 8] where the following bound is proved

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

The theorem is proved plugging the above estimate into (23). □

We add the following remark.

Remark 4 Using the Hoeffding inequality one can show that for $|f(x)| < C$ the following estimate holds true:

$$\mathbb{P}\left(\left| \|f\|_{\rho} - \|f\|_{\rho_{\mathbf{z}}} \right| > \epsilon \right) \leq 2e^{-\frac{n\epsilon^2 \|f\|_{\rho}^2}{2C^4}},$$

that is we have with probability at least $1 - \eta$,

$$\left| \|f\|_{\rho} - \|f\|_{\rho_{\mathbf{z}}} \right| \leq \sqrt{2} C^2 \left(\frac{\log(2/\eta)}{n} \right)^{\frac{1}{2}} \|f\|_{\rho}^{-1}. \tag{24}$$

Comparing the above result to Proposition 1 we see that one has the order $n^{-1/2}$ versus $n^{-1/4}$. It is hence tempting to use this estimate instead of that in Proposition 1 to avoid dealing with RKHS. The point is that we need to estimate $|\|f\|_{\rho} - \|f\|_{\rho_{\mathbf{z}}}|$ for $f = f_{\mathcal{H}} - f_{\mathbf{z}}^{\lambda}$, and it is expected that the norm $\|f\|_{\rho}$ is rather small, say

$$\|f\|_{\rho} = \left\| f_{\mathcal{H}} - f_{\mathbf{z}}^{\lambda} \right\|_{\rho} \leq c n^{-\frac{r}{2r+1}}, \quad r > \frac{1}{2},$$

as in Example 1. Note that for such f the bound (24) is too rough. Namely,

$$\sqrt{2} C^2 \left(\frac{\log(2/\eta)}{n} \right)^{\frac{1}{2}} \|f\|_{\rho}^{-1} = \mathcal{O}(n^{-\frac{1}{2(2r+1)}}) \gg n^{-\frac{1}{4}}.$$

The simple application of the Hoeffding inequality is then not enough to prove optimal learning rates and in the sequel we will use the bound given in Proposition 1. We are grateful to an anonymous referee who inspired us to make this remark.

Assumption 2 and Proposition 1 immediately yield 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 substituting (15), (16) into the above inequality and noting that $n^{-1/4} \leq \sqrt{\lambda}$ since $\lambda \geq n^{-1/2}$. □

Given the above results we can prove Theorem 3.

Proof of Theorem 3 We first note a few useful facts. Let $\Theta(\lambda) = \omega(\lambda)\mathcal{A}(\lambda)$. First, from Assumption 2, item 3, if we take $\lambda = \lambda_{\mathrm{o}}(n)$ we have

$$\Theta\big(\lambda_{\mathrm{o}}(n)\big) \leq c\lambda_{\mathrm{o}}(n) \quad \Rightarrow \quad \frac{1}{\sqrt{n}} \leq c\lambda_{\mathrm{o}}(n) \quad \Rightarrow \quad \frac{1}{n^{1/4}} \leq \sqrt{c}\sqrt{\lambda_{\mathrm{o}}(n)}. \tag{25}$$

Second, noting that (19) 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

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

This immediately yields

$$\frac{1}{\omega(\lambda_{\rho_{\mathbf{z}}})} \leq \frac{q}{\omega(\lambda_{\mathrm{o}}(n))}, \tag{26}$$

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

$$\frac{\sqrt{\lambda_{\mathcal{H}}}}{\omega(\lambda_{\mathcal{H}})} \leq \frac{q\sqrt{\lambda_{\mathrm{o}}(n)}}{\omega(\lambda_{\mathrm{o}}(n))}, \tag{27}$$

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

$$\begin{aligned}
\left\| f_{\mathbf{z}}^{\hat{\lambda}} - f_{\mathcal{H}} \right\|_{\rho} &\leq \left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathcal{H}} \right\|_{\rho_{\mathbf{z}}} + \frac{\alpha(\eta) C_{\kappa}}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathcal{H}} \right\|_{\mathcal{H}} \\
&\leq \left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathcal{H}} \right\|_{\rho_{\mathbf{z}}} + \frac{\alpha(\eta) C_{\kappa}}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} \right\|_{\mathcal{H}} \\
&\quad + \frac{\alpha(\eta) C_{\kappa}}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathcal{H}} \right\|_{\mathcal{H}}.
\end{aligned} \tag{28}$$

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

$$\left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathcal{H}} \right\|_{\rho_{\mathbf{z}}} \leq 6q\alpha(\eta)\hat{C}\sqrt{\lambda_{\mathrm{o}}(n)}\mathcal{A}\big(\lambda_{\mathrm{o}}(n)\big). \tag{29}$$

Applying again Theorem 1 and with the aid of (25) we obtain

$$\frac{\alpha(\eta) C_{\kappa}}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathcal{H}} \right\|_{\mathcal{H}} \leq 6q\alpha(\eta)^2 c C_{\kappa} \sqrt{\lambda_{\mathrm{o}}(n)}\mathcal{A}\big(\lambda_{\mathrm{o}}(n)\big). \tag{30}$$

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

$$\left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} \right\|_{\mathcal{H}} \leq \frac{4\alpha(\eta)}{\sqrt{n}\omega(\lambda_{\rho_{\mathbf{z}}})}. \tag{31}$$

We can now use (25), (26) and the definition of $\lambda_{\mathrm{o}}(n)$ to get

$$\frac{\alpha(\eta)C_{\kappa}}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} \right\|_{\mathcal{H}} \leq 4q\alpha(\eta)^2 c C_{\kappa} \sqrt{\lambda_{\mathrm{o}}(n)} \mathcal{A}\big(\lambda_{\mathrm{o}}(n)\big). \tag{32}$$

If we now substitute (29), (30), (32) into (28) we get

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

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

$$\begin{aligned} \left\| f_{\mathbf{z}}^{\hat{\lambda}} - f_{\mathcal{H}} \right\|_{\rho} &\leq \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathcal{H}} \right\|_{\rho_{\mathbf{z}}} + \frac{\alpha(\eta)C_{\kappa}}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathcal{H}} \right\|_{\mathcal{H}} \\ &\leq \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} \right\|_{\rho_{\mathbf{z}}} + \left\| f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} - f_{\mathcal{H}} \right\|_{\rho_{\mathbf{z}}} \\ &\quad + \frac{\alpha(\eta)C_{\kappa}}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathcal{H}} \right\|_{\mathcal{H}}. \end{aligned} \tag{33}$$

Applying Theorem 1 and using (25) we immediately get

$$\frac{\alpha(\eta)C_{\kappa}}{n^{1/4}} \left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathcal{H}} \right\|_{\mathcal{H}} \leq 6qc\alpha(\eta)^2 C_{\kappa} \sqrt{\lambda_{\mathrm{o}}(n)} \mathcal{A}\big(\lambda_{\mathrm{o}}(n)\big). \tag{34}$$

Another straightforward application of Theorem 1 and Corollary 1 gives

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

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

$$\left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} \right\|_{\rho_{\mathbf{z}}} \leq \frac{4\alpha(\eta)\hat{C}\sqrt{\lambda_{\mathcal{H}}}}{\sqrt{n}\omega(\lambda_{\mathcal{H}})}, \tag{36}$$

so that using (27), (25) and the definition of $\lambda_{\mathrm{o}}(n)$ we can write

$$\left\| f_{\mathbf{z}}^{\lambda_{\mathcal{H}}} - f_{\mathbf{z}}^{\lambda_{\rho_{\mathbf{z}}}} \right\|_{\rho_{\mathbf{z}}} \leq 4\alpha(\eta) q \hat{C} \sqrt{\lambda_{\mathrm{o}}(n)} \mathcal{A}\big(\lambda_{\mathrm{o}}(n)\big). \tag{37}$$

The proof is finished by substituting (34), (35) and (37) into (33) to get

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

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 a very small sample. When the number of samples is very small, as is often the case in practice, we observe that one cannot completely rely on the theoretical constructions since the bounds 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 the 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 given a priori. 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

$$K(\lambda) = \arg\min\{Q_{\mathbf{z}}(K,\lambda),\ K \in \mathbb{K}\}, \tag{38}$$

where

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

and $\mathcal{H}_K$ is the RKH space generated by K. By definition, the selected kernel $K = K(\lambda)$ is λ-dependent, so that this kernel choice rule is only applicable for an 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 Theorem 3 the best parameter choice rule $\lambda = \hat{\lambda} = \hat{\lambda}(K)$ allows for an accuracy which is only by a constant factor worse than the optimal one for fixed $K \in \mathbb{K}$.

Let $\Lambda\colon \mathbb{R}_+ \to \mathbb{R}_+$ be the function such that its value at point λ is the best parameter choice $\hat{\lambda} = \hat{\lambda}(K(\lambda))$ for estimators based on the kernel $K(\lambda) \in \mathbb{K}$ given by (38). If $\mathring{\lambda}$ is a fixed 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 a fixed point $\lambda = \mathring{\lambda}$ depends on the set $\mathbb{K}$, and deserves consideration in the future. In the computational tests below we find such a fixed point numerically for an academic example from [29]. At this point it is worth noting that the balancing principle can be capacity independent in the sense that it does not require a knowledge of the spectral properties of the underlying kernel K. This feature of the balancing principle makes its combination with the rule (38) numerically feasible.

To simplify the numerical realization of the balancing principle and especially in the presence of very small samples, one can approximate the values $\lambda_{\rho_{\mathbf{z}}}$, $\lambda_{\mathcal{H}}$ using the well-known quasi-optimality criterion [37]. As observed in [30] this criterion can

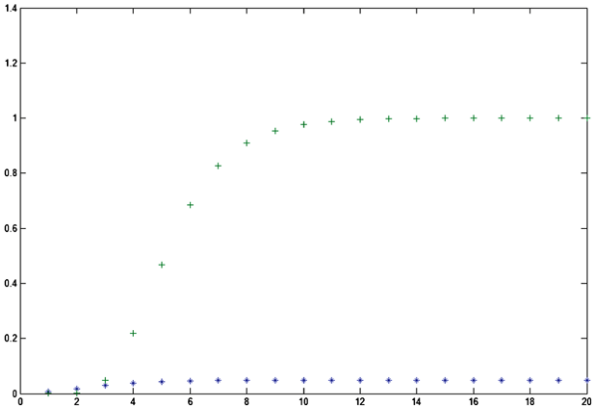

Fig. 2 (Color online) The values of $\sigma_{\rho_{\mathbf{z}}}(j)$ (*blue dots*) and $\sigma_{\mathcal{H}}(j)$ (*green crosses*) for $\mathbf{z}=\mathbf{z}_{21}$

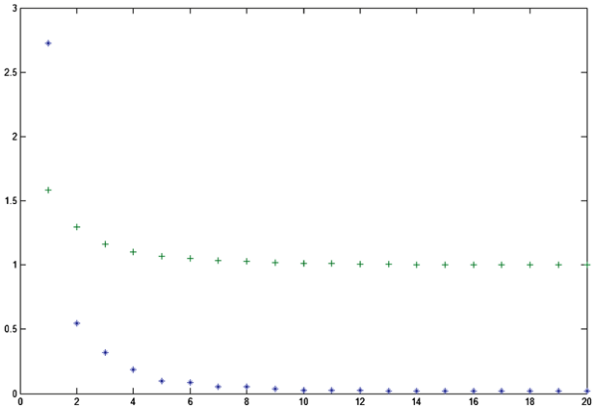

Fig. 3 (Color online) The values of $\sigma_{\rho_{\mathbf{z}}}(j)$ (*blue dots*) and $\sigma_{\mathcal{H}}(j)$ (*green crosses*) for $\mathbf{z}=\mathbf{z}_{51}$

be seen as a heuristic counterpart of the parameter choice rule $\lambda=\bar{\lambda}$. It also uses the norms $\sigma(j)=\|f_{\mathbf{z}}^{\lambda_j}-f_{\mathbf{z}}^{\lambda_{j-1}}\|$, $\lambda_j=\lambda_{\text{start}}\cdot\mu^j$, and selects $\lambda^{q-0}=\lambda_l$ such that for any $j=1,2,\ldots,N$, $\sigma(j)\geq\sigma(l)$, i.e.,

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

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

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

and

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

respectively. Then in accordance with (14) we choose a regularization parameter

$$\hat{\lambda}=\min\bigl\{\lambda_{\rho_{\mathbf{z}}}^{q-0},\lambda_{\mathcal{H}}^{q-0}\bigr\}. \tag{39}$$

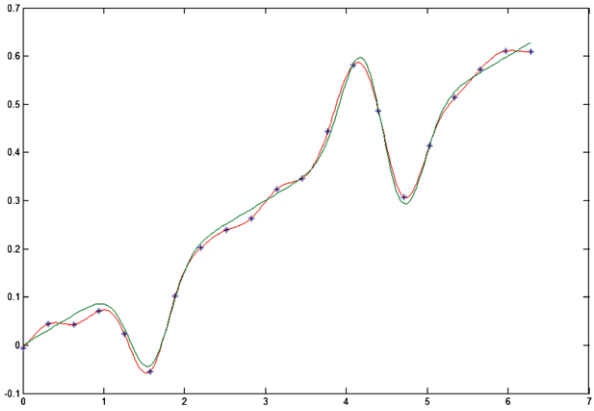

Fig. 4 (Color online) The estimator $f_{\mathbf{z}}^{\hat{\lambda}}$ (*red line*) and the target function f_ρ (*green line*) for $\hat{\lambda} = 1.5 \times 10^{-6}$ and training set $\mathbf{z} = \mathbf{z}_{21}$ (*blue dots*)

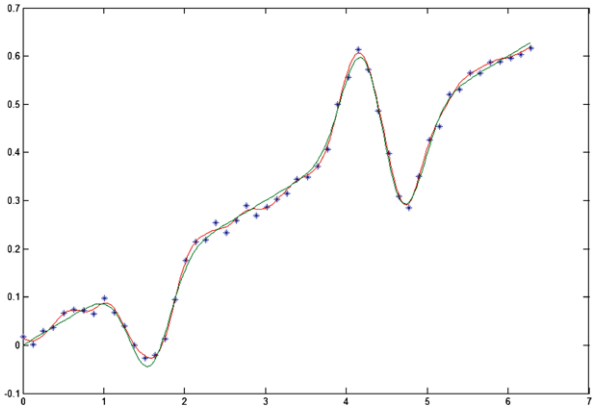

Fig. 5 (Color online) The estimator $f_{\mathbf{z}}^{\hat{\lambda}}$ (*red line*) and the target function f_ρ (*green line*) for $\hat{\lambda} = 0.0033$ and training set $\mathbf{z} = \mathbf{z}_{51}$ (*blue dots*)

As in [29], we consider a target function

$$f_\rho(x) = \frac{1}{10}\big(x + 2\big(\mathrm{e}^{-8(\frac{4}{3}\pi - x)^2} - \mathrm{e}^{-8(\frac{\pi}{2} - x)^2} - \mathrm{e}^{-8(\frac{3}{2}\pi - x)^2}\big)\big), \quad x \in [0, 2\pi], \tag{40}$$

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 the approximate version (39) of the balancing principle using the a priori information that the target function (40) belongs to a RKH space $\mathcal{H} = \mathcal{H}_K$ generated by the kernel $K(x,t) = K_\rho(x,t) = xt + \mathrm{e}^{-8(t-x)^2}$, $t, x \in [0, 2\pi]$.

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

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

It is instructive to see that the sequences $\sigma_{\rho_{\mathbf{z}}}(j)$, $\sigma_{\mathcal{H}}(j)$, $j = 1, 2, \ldots, 20$, exhibit different behavior for the training sets $\mathbf{z}_{21}$ and $\mathbf{z}_{51}$. At the same time, they attain their

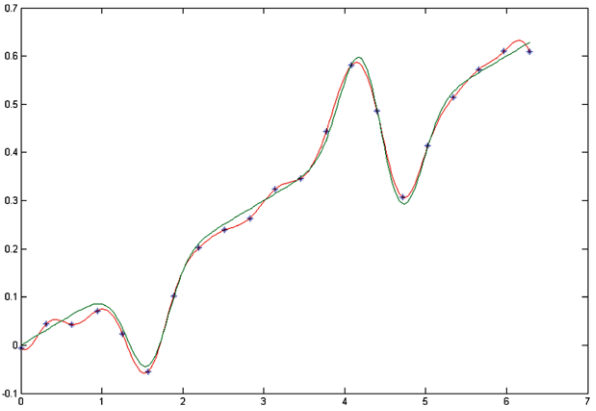

Fig. 6 (Color online) The target function f_ρ (*green line*) and its estimator $f_{\mathbf{z}}^{\mathring{\lambda}}$ (*red line*) based on the adaptively chosen kernel $K(\mathring{\lambda}; x, t) = xt + \mathrm{e}^{-10(x-t)^2}$, $\mathring{\lambda} = 0.0014$, and training set $\mathbf{z} = \mathbf{z}_{21}$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

minimal values at the same j. Therefore, in accordance with the rule (39) we take $\hat{\lambda} = \lambda_{\rho_{\mathbf{z}}}^{q-0} = \lambda_{\mathcal{H}}^{q-0} = 1.5 \times 10^{-6}$ in the case of $\mathbf{z} = \mathbf{z}_{21}$, while for $\mathbf{z} = \mathbf{z}_{51}$ $\hat{\lambda} = \lambda_{\rho_{\mathbf{z}}}^{q-0} = \lambda_{\mathcal{H}}^{q-0} = 0.0033$.

Figures 4 and 5 show that for the chosen values of the 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 (40). Instead, we choose a kernel K adaptively from the set

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

trying to find a fixed point of the function $\Lambda\colon \lambda \to \hat{\lambda}(K(\lambda))$, where $\hat{\lambda}(K(\lambda))$ is the number (39) calculated for the kernel $K(\lambda)$, which minimizes $Q_{\mathbf{z}}(K, \lambda)$ for $\mathbf{z} = \mathbf{z}_{21}$, 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 a simple complete search over the finite set $\mathbb{K}$. Then the next value $\lambda^{(s+1)} \in \{\lambda_j\}_{j=1}^{20}$ is defined as the number (39) calculated for the estimators $f_{\mathbf{z}}^{\lambda_j}$ based on the kernel $K(\lambda^{(s)})$. This iterative procedure terminates when $|\lambda^{(s+1)} - \lambda^{(s)}| \le 10^{-4}$. It gives us the required approximate fixed point $\mathring{\lambda} = \lambda_{18} \approx 0.0014$ and the corresponding kernel $K(\mathring{\lambda}) = K(\mathring{\lambda}; x, t) = xt + \mathrm{e}^{-10(x-t)^2}$, 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 (40), as can be seen in Fig. 6.

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

Acknowledgements The authors thank the anonymous reviewers for many useful comments and a careful review of the paper. 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 7th Framework Programme of EC. Ernesto De Vito and Lorenzo Rosasco have been partially supported by the FIRB project RBIN04PARL and by 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.

References

1. N. Aronszajn, Theory of reproducing kernels, *Trans. Am. Math. Soc.* **68**, 337–404 (1950).
2. A. Barron, L. Birgé, P. Massart, Risk bounds for model selection via penalization, *Probab. Theory Relat. Fields* **113**(3), 301–413 (1999).
3. P.L. Bartlett, S. Boucheron, G. Lugosi, Model selection and error estimation, in Proceedings of the Thirteenth Annual Conference on Computational Learning Theory (2000), pp. 286–297.
4. F. Bauer, S. Pereverzev, L. Rosasco, On regularization algorithms in learning theory, *J. Complex.* **23**(1), 52–72 (2007).
5. S. Boucheron, O. Bousquet, G. Lugosi, Theory of classification: a survey of some recent advances, *ESAIM Probab. Stat.* **9**, 323–375 (2005) (electronic).
6. P. Bühlmann, B. Yu, Boosting with the l_2-loss: Regression and classification, *J. Am. Stat. Assoc.* **98**, 324–340 (2002).
7. A. Caponnetto, Optimal rates for regularization operators in learning theory, Technical report, CBCL Paper #264/ CSAIL-TR #2006-062, MIT (2006). Available at http://cbcl.mit.edu/projects/cbcl/publications/ps/MIT-CSAIL-TR-2006-062.pdf.
8. A. Caponnetto, E. De Vito, Optimal rates for the regularized least-squares algorithm, *Found. Comput. Math.* **7**(3), 331–368 (2007).
9. A. Caponnetto, 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).
10. A. Christmann, I. Steinwart, Consistency and robustness of kernel-based regression in convex risk minimization, *Bernoulli* **13**(3), 799–819 (2007).
11. F. Cucker, S. Smale, On the mathematical foundations of learning, *Bull. Am. Math. Soc. (NS)* **39**(1), 1–49 (2002) (electronic).
12. F. Cucker, D.-X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Monographs on Applied and Computational Mathematics (Cambridge University Press, Cambridge, 2007). With a foreword by S. Smale.
13. C. De Mol, E. De Vito, L. Rosasco, Elastic-net regularization in learning theory, *J. Complex.* **25**(2), 201–230 (2009).
14. E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, F. Odone, Learning from examples as an inverse problem, *J. Mach. Learn. Res.* **6**, 883–904 (2005).
15. E. De Vito, L. Rosasco, A. Verri, Spectral methods for regularization in learning theory, Technical Rerport DISI-TR-05-18, DISI, Universitá degli Studi di Genova, Italy (2005).
16. R. DeVore, G. Kerkyacharian, D. Picard, V. Temlyakov, On mathematical methods of learning, *Found. Comput. Math.* **6**(1), 3–58 (2006).
17. L. Devroye, L. Györfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Applications of Mathematics, vol. 31 (Springer, New York, 1996).
18. S. Dudoit, M. van der Laan, Asymptotics of cross-validated risk estimation in estimator selection and performance assessment, *Stat. Methodol.* **2**(2), 131–154 (2005).
19. H.W. Engl, M. Hanke, A. Neubauer, *Regularization of inverse problems*, Mathematics and Its Applications, vol. 375 (Kluwer Academic, Dordrecht, 1996).
20. T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, *Adv. Comput. Math.* **13**, 1–50 (2000).
21. S. Gaïffas, G. Lecué, Aggregation of penalized empirical risk minimizers in regression, Preprint (2009).
22. A. Goldenshluger, S. Pereverzev, On adaptive inverse estimation of linear functionals in Hilbert scales, *Bernoulli* **9**(5), 783–807 (2003).
23. L. Györfi, M. Kohler, A. Krzyzak, H. Walk, *A Distribution-Free Theory of Non-Parametric Regression*, Springer Series in Statistics (Springer, New York, 2002).
24. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2001).

25. V. Koltchisnkii, Local Rademacher complexities and oracle inequalities in risk minimization, *Ann. Stat.* **34**(6), 2593–2656 (2004).
26. O. Lepskii, On a problem of adaptive estimation in Gaussian white noise, *Theory Probab. Its Appl.* **35**, 454–466 (1990).
27. G. Lugosi, M. Wegkamp, Complexity regularization via localized random penalties, *Ann. Statist.* **32**(4), 1679–1697 (2004).
28. P. Mathé, The Lepskiĭ principle revisited, *Inverse Probl.* **22**(3), L11–L15 (2006).
29. C. Micchelli, M. Pontil, Learning the kernel function via regularization, *J. Mach. Learn. Res.* **6**, 1099–1125 (2005) (electronic).
30. S. Pereverzev, E. Schock, On the adaptive selection of the parameter in regularization of ill-posed problems, *SIAM J. Numer. Anal.* **43**(5), 2060–2076 (2005) (electronic).
31. T. Poggio, F. Girosi, A theory of networks for learning, *Science* **247**, 978–982 (1990).
32. L. Rosasco, *Regularization Approaches in Learning Theory*, PhD Thesis, University of Genova (2006).
33. R. Rosipal, L.J. Trejo, A. Cichocki, Kernel principal component regression with an EM approach to nonlinear principal components extraction, Technical report, University of Paisley (2000).
34. B. Schölkopf, A.J. Smola, *Learning with Kernels* (MIT Press, Cambridge, 2002).
35. S. Smale, D.-X. Zhou, Learning theory estimates via integral operators and their approximations, *Constr. Approx.* **26**(2), 153–172 (2007).
36. I. Steinwart, A. Christmann, *Support Vector Machines*, Information Science and Statistics (Springer, New York, 2008).
37. A.N. Tikhonov, V.B. Glasko, Use of the regularization method in non-linear problems, *Zh. Vychisl. Mat. Mat. Fiz.* **5**, 463–473 (1965).
38. A. Tsybakov, *Introduction to Nonparametric Estimation*, Springer Series in Statistics (Springer, Berlin, 2008).
39. A.B. Tsybakov, Optimal aggregation of classifiers in statistical learning, *Ann. Stat.* **32**, 135–166 (2004).
40. A. van der Vaart, S. Dudoit, M. van der Laan, Oracle inequalities for multi-fold cross validation. *Stat. Decis.* 24(3), 2006.
41. V.N. Vapnik, *Statistical Learning Theory, Adaptive and Learning Systems for Signal Processing, Communications, and Control* (Wiley, New York, 1998).
42. G. Wahba, *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59 (SIAM, Philadelphia, 1990).
43. Q. Wu, Y. Ying, D.-X. Zhou, Learning rates of least-square regularized regression, *Found. Comput. Math.* **6**(2), 171–192 (2006).
44. Y. Yao, L. Rosasco, A. Caponnetto, On early stopping in gradient descent learning, *Constr. Approx.* **26**(2), 289–315 (2007).
45. H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *JRSSB* **67**(2), 301–320 (2005).