Johann Radon Institute for
Computational and Applied Mathematics
Austrian Academy of Sciences (ÖAW)

A Linear Functional Strategy for Regularized Ranking

G. Kriukova, O. Panasiuk, S. Pereverzyev,
P. Tkachenko

RICAM-Report 2015-13

A Linear Functional Strategy for Regularized Ranking

Galyna Kriukova, Oleksandra Panasiuk, Sergei V. Pereverzyev, Pavlo Tkachenko

Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstrasse 69, 4040 Linz, Austria

Abstract

Regularization schemes are frequently used for performing ranking tasks. This topic has been intensively studied in recent years. However, to be effective a regularization scheme should be equipped with a suitable strategy for choosing a regularization parameter. In the present study we discuss an approach, which is based on the idea of a linear combination of regularized rankers corresponding to different values of the regularization parameter. The coefficients of the linear combination are estimated by means of the so-called linear functional strategy. We provide a theoretical justification of the proposed approach and illustrate them by numerical experiments. Some of them are related with ranking the risk of nocturnal hypoglycemia of diabetes patients.

Keywords: Regularization, Ill-posed problem, Ranking, Linear functional strategy, Diabetes technology

1. Introduction

In supervised learning one is often given a sequence of examples of $x = x_1, x_2, \ldots, x_m \in X \subset \mathbb{R}^d$ labeled with the corresponding values $y_1, y_2, \ldots, y_m \in Y \subset \mathbb{R}$ of the dependent variable y. Then the learning task is to use this data as a training set $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^m \subset Z = X \times Y$ for assigning a proper label y to a previously unseen $x \in X$.

If the labels y_i are treated as the values of some function at given points x_i, then the above mentioned learning task is referred to as regression, or regression learning, and is one of the most well-studied problems in learning theory. In recent years another problem called ranking has gained attention in this theory.

Ranking is relatively new learning problem that is parallel to regression. After the first paper [1] was published in 1999, ranking has been intensively investigated in the literature. Here we refer to [2, 3, 4, 5, 6, 7, 8], just to mention a few publications.

In ranking one also learns a real-valued function $f : X \to Y$ that assigns a label y to $x \in X$, but the value $y = f(x)$ itself is not so important. What do matter are the relative ranks of instances $x, x' \in X$ induced by the labels $f(x), f(x')$. Namely, a ranking function $f : X \to Y$ ranks instances x with larger labels $f(x)$ higher than those with smaller labels.

Thus, the task of learning ranking is different from regression, but if we are looking for labeling functions $f : X \to Y$ in some Reproducing Kernel Hilbert Space (RKHS) on X then, in spite of the difference, both learning problems can be formulated as ill-posed linear integral operator equations of the first kind in the chosen RKHS [9, 8, 10]. The ill-posedness of such formulations calls for the employment of the regularization theory in the construction of regression and ranking algorithms. In this theory, the performance of algorithms is usually estimated under the source conditions expressed in terms of the so-called index functions. It is known (see, e.g. [11, 12]) that ill-posed equations may involve entirely different operators but nevertheless allow the same performance of regularization algorithms, if the solutions of these equations satisfy the source conditions for the same index function.

On the other hand, recently, the authors of [13] have noticed a suboptimality of known ranking performance estimates compared to the corresponding regression ones. The same observation can be made from the comparison of [8] with [9] and [14] with [15], when the same regularization schemes are compared under the source conditions generated by the same index functions.

This observation is not in agreement with the general fact of the regularization theory mentioned above, and it hints at a gap in the analysis of the regularized ranking algorithms. In the present paper we refine this analysis and show that, at least for the so-called offline learning, the performance of the regularized ranking is similar to that of the regularized regression learning.

The above mentioned refinement is obtained as a by-product of the study of a new a posteriori regularization scheme in the context of learning. Note that usually a posteriori regularization means an adaptive choice of the parameter for single-parameter regularization methods such as Tikhonov, Lavrentiev or Landweber regularization and others like that. In the existing literature on the regularization theory it is suggested to make the above

choice by using one of the known rules such as quasi-optimality criterion, cross-validation, the discrepancy principle, the balancing principle. In the context of learning these rules have been discussed in [16, 17]. But these and similar rules select only one element from a family of approximants, calculated according to an employed regularization method, and leave others aside. Of course, the other approximants are used in the selection process, but then they are rejected, in spite of the numerical expenses made for their construction. At the same time, the rejected approximants may also contain important information on the approximated quantity of interest and can contribute to the improvement of the accuracy of its reconstruction (see Figure 1 below).

In the present study we explore the idea to use the calculated approximants in the construction of a new one. More precisely, the idea is to use linear combinations of the approximants calculated for different values of the regularization parameter. It is clear that the best Hilbert-space approximation by such a linear combination requires the knowledge of inner products between the calculated approximants and the approximated element, which is of course unknown.

At the same time, the regularization theory tells us (see, e.g. [11], Proposition 2.17) that the values of linear bounded functionals (e.g., inner products) at the approximated elements can be estimated more accurately than the elements themselves. The idea is to use the estimated values of the corresponding inner products for mimicking the best linear combination of the calculated regularized approximants.

In the regularization theory the above-mentioned accurate estimation of linear functionals is often called as linear functional strategy (LFS). It was proposed in [18] and then further developed in [19, 20, 21]. The previous results on LFS have been obtained under the assumption that the operators from the considered ill-posed equations are directly accessible, but that is not the case in the learning context. Therefore in the present study we at first perform adaptation-extension of LFS to that context.

The paper is organized as follows. In the next section we recall the setting of least squares ranking and its formulation as an ill-posed linear operator equation. Moreover, we describe a general regularization scheme for solving this equation. At the end of the section we specify the idea of a linear combination of regularized rankers. In Section 3 we present the above-mentioned extension of LFS and use it for mimicking the best approximation by linear combinations of given rankers. The section also contains new bounds on the

excess risk of the regularized ranking. In Section 4 we illustrate our theoretical results by numerical tests and discuss an application of the proposed ranking scheme in diabetes technology.

2. Problem Setting

Let the inputs x be taken from a compact domain or a manifold X in the Euclidean space $\mathbb{R}^d$ and the ranking output space is $Y = [-M, M] \subset \mathbb{R}$. The input x and the output y are assumed to be related by a conditional probability distribution $\rho(y|x)$ of y given x. Moreover, the input x is also assumed to be random and governed by an unknown marginal probability ρ_X on X so that there is an unknown probability distribution $\rho(x, y) = \rho_X(x)\rho(y|x)$ on the sample space $Z = X \times Y$ from which the data forming the training set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ are drawn independently. We are interested in synthesizing a function $y = f(x)$ that will mimic the relation between the inputs x and the corresponding outputs y. More precisely, the ranking problem is to learn from $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ a ranking function $f = f_{\mathbf{z}} : X \to Y$.

For given true ranks y and y' of the inputs $x, x' \in X$ the value

$$(y - y' - (f(x) - f(x')))^2$$

is interpreted as the *magnitude-preserving least squares loss* of a ranking function f (see [7, 8, 13, 22]). Then the quality of a ranking function f can be measured by the expected risk

$$\mathcal{E}(f) = \int_Z \int_Z (y - y' - (f(x) - f(x')))^2 \, d\rho(x, y) d\rho(x', y')$$

Let $\mathcal{F}_\rho$ be a set of functions minimizing the risk $\mathcal{E}(f)$. As it has been noticed in [8, 13], $\mathcal{F}_\rho$ contains the target function

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

also known in learning theory as the *regression function*. It is easy to observe, that the target function is not unique, for instance $f_\rho(x) + c \in \mathcal{F}_\rho$ for each $c \in \mathbb{R}$.

The ideal estimator $f_\rho(x)$ can not be found in practice, because the conditional probability distribution $\rho(y|x)$ is unknown. Therefore, the goal might be to find f minimizing the excess risk $\mathcal{E}(f) - \mathcal{E}(f_\rho)$ over some hypothesis

space $\mathcal{H} \in L_2(X, \rho_X)$. A widely used choice of such a space is a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H} = \mathcal{H}_K$, associated with a kernel $K : X \times X \to \mathbb{R}$.

Observe that from the very definition of $\mathcal{E}(f)$ and f_ρ it follows that

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \int_{X\times X} \left[(f(x) - f(x')) - (f_\rho(x) - f_\rho(x'))\right]^2 d\rho_X(x) d\rho_X(x') \quad (1)$$

Indeed,

$$\begin{aligned}
\mathcal{E}(f) - \mathcal{E}(f_\rho) &= \int_{Z\times Z} \left(y - y' - (f(x) - f(x'))\right)^2 d\rho(x, y) d\rho(x', y') \\
&\quad - \int_{Z\times Z} \left(y - y' - (f_\rho(x) - f_\rho(x'))\right)^2 d\rho(x, y) d\rho(x', y') \\
&= \int_X \int_Y \int_X \int_Y [2y - 2y' - f_\rho(x) + f_\rho(x') - f(x) + f(x')] \\
&\quad \times [(f_\rho(x) - f_\rho(x')) - (f(x) - f(x'))] d\rho(y'|x') d\rho(y|x) d\rho_X(x') d\rho_X(x) \\
&= \int_{X\times X} [(f_\rho(x) - f_\rho(x')) - (f(x) - f(x'))]^2 d\rho_X(x) d\rho_X(x').
\end{aligned}$$

Consider the space $L_2(X^2, \rho_{X^2})$ of square-integrable functions $g(x, x')$ with respect to the product measure $d\rho_{X^2}(x, x') = d\rho_X(x) d\rho_X(x')$ on $X^2 = X \times X$, and the operators $\Delta : L_2(X, \rho_X) \to L_2(X^2, \rho_{X^2})$, $\mathcal{D}_K : \mathcal{H}_K \to L_2(X^2, \rho_{X^2})$ such that

$$\begin{aligned}
\Delta f(x, x') &= f(x) - f(x'), \\
\mathcal{D}_K f(x, x') &= \langle K_x - K_{x'}, f\rangle_{\mathcal{H}_K} = f(x) - f(x'),
\end{aligned}$$

where $K_x = K(x, \cdot)$, $K_{x'} = K(x', \cdot)$, and we use the reproducing property $f(x) = \langle K_x, f\rangle_{\mathcal{H}_K}$. Then in view of (1) the minimization of the excess risk $\mathcal{E}(f) - \mathcal{E}(f_\rho)$ can be written as the least squares problem

$$\|\mathcal{D}_K f - \Delta f_\rho\|^2_{L_2(X^2, \rho_{X^2})} \to \min$$

that leads to the normal equation

$$\mathcal{D}_K^* \mathcal{D}_K f = \mathcal{D}_K^* \Delta f_\rho, \quad (2)$$

where $\mathcal{D}_K^*; L_2(X^2, \rho_{X^2}) \to \mathcal{H}_K$ is the adjoint of $\mathcal{D}_K$ and defined as follows:

$$\mathcal{D}_K^* g(\cdot) = \int_{X\times X} [K(x, \cdot) - K(x', \cdot)] g(x, x') d\rho_X(x) d\rho_X(x').$$

With the use of the operator

$$L_K f(\cdot) = \int_{X\times X} [K(x,\cdot) - K(x',\cdot)] f(x) d\rho_X(x) d\rho_X(x'),$$

which has been studied in a slightly different context in [8, 14], we can represent $\mathcal{D}_K^* \mathcal{D}_K$ as

$$\begin{aligned}\mathcal{D}_K^* \mathcal{D}_K f(\cdot) &= \int_{X\times X} [K(x,\cdot) - K(x',\cdot)](f(x) - f(x')) d\rho_X(x) d\rho_X(x') \\ &= \int_{X\times X} [K(x,\cdot) - K(x',\cdot)] f(x) d\rho_X(x) d\rho_X(x') \\ &\quad - \int_{X\times X} [K(x,\cdot) - K(x',\cdot)] f(x') d\rho_X(x) d\rho_X(x') = 2L_K f(\cdot).\end{aligned}$$

Then it is clear that $L_K : \mathcal{H}_K \to \mathcal{H}_K$ is a self-adjoint and positive operator. Moreover, its definition domain can be naturally extended to $L_2(X, \rho_X)$, and in the same way as above we have $\mathcal{D}_K^* \Delta f(\cdot) = 2L_K f(\cdot)$, but $L_K : L_2(X, \rho_X) \to \mathcal{H}_K \subset L_2(X, \rho_X)$ is not a self-adjoint operator. Nevertheless, with a slight abuse of notations, one may rewrite the equation (2) in the form

$$L_K f = L_K f_\rho \tag{3}$$

and look for its solution f in the hypothesis space $\mathcal{H}_K$.

At the same time, one should be aware of the fact that, in general, the equation (3) is ill-posed. Therefore, instead of minimizing the excess risk $\mathcal{E}(f) - \mathcal{E}(f_\rho)$ one minimizes its regularized version

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) + \alpha \|f\|^2_{\mathcal{H}_K} \to \min \tag{4}$$

that accordingly leads to Lavrentiev regularization of the equation (3)

$$\frac{\alpha}{2} f + L_K f = L_K f_\rho, \tag{5}$$

and the latter one has been discussed in [8].

At this point it should be noted that neither (3), nor (5) is accessible because the distribution ρ_X is not known. This forces to minimize an empirical version of (4)

$$\frac{1}{m^2} \sum_{i,j=1}^{m} (y_i - y_j - (f(x_i) - f(x_j)))^2 + \alpha \|f\|^2_{\mathcal{H}_K} \to \min,$$

that in its turn leads to the discretized version of (5)

$$\frac{\alpha}{2}f + \frac{1}{m^2}S_{\mathbf{x}}^*\mathbb{D}S_{\mathbf{x}}f = \frac{1}{m^2}S_{\mathbf{x}}^*\mathbb{D}\mathbf{y}, \tag{6}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_m)^T \in \mathbb{R}^m$, $\mathbb{D} = m\mathbb{I} - \mathbf{1} \times \mathbf{1}^T$, and $\mathbb{I}$, $\mathbf{1}$ are the m-th order unit matrix and the vector of all ones, $S_{\mathbf{x}} : \mathcal{H}_K \to \mathbb{R}^m$ is the sampling operator $S_{\mathbf{x}}f = (f(x_1), f(x_2), \ldots, f(x_m))^T$ associated with a discrete set $\mathbf{x} = \{x_i\}_{i=1}^m \subset X$, and $S_{\mathbf{x}}^* : \mathbb{R}^m \to \mathcal{H}_K$ is the adjoint of $S_{\mathbf{x}}$.

The comparison of (5) and(6) allows the conclusion that $\frac{1}{m^2}S_{\mathbf{x}}^*\mathbb{D}S_{\mathbf{x}}$ and $\frac{1}{m^2}S_{\mathbf{x}}^*\mathbb{D}\mathbf{y}$ can be used as available approximations of L_K and $L_K f_\rho$ respectively.

2.1. One-parameter regularization

The whole arsenal of regularization methods can potentially be used to approximately solve the initial equation (3). In particular, one can use the general single-parameter regularization scheme and construct an approximate solution $f_{\mathbf{z}} = f_{\mathbf{z}}^\alpha$ to (3) as

$$f_{\mathbf{z}}^\alpha = g_\alpha(\tfrac{1}{m^2}S_{\mathbf{x}}^*\mathbb{D}S_{\mathbf{x}})\tfrac{1}{m^2}S_{\mathbf{x}}^*\mathbb{D}\mathbf{y},$$

where $\{g_\alpha\}$ is a *one-parameter regularization family.* Let us recall its definition:

Definition 1. *A family of functions $g_\alpha : [0, d] \to \mathbb{R}$ parametrized by a parameter $\alpha > 0$ is called a regularization on $[0, d]$, if there are constants γ_{-1}, γ_0 for which*

$$\sup_{0<t\le d} |1 - tg_\alpha(t)| \le \gamma_0, \tag{7}$$

$$\sup_{0<t\le d} |g_\alpha(t)| \le \frac{\gamma_{-1}}{\alpha}. \tag{8}$$

Definition 2. *A regularization scheme generated by a family of functions $\{g_\alpha\}$ has a qualification p if there is a constant $\gamma_p > 0$ such that for any $\alpha > 0$ if holds*

$$\sup_{0<t\le d} t^p|1 - tg_\alpha(t)| \le \gamma_p\alpha^p.$$

Remark 1. *Lavrentiev regularization is generated by*

$$g_\alpha(t) = \left(\tfrac{\alpha}{2} + t\right)^{-1},$$

and has qualification $p = 1$.

p-times iterated Lavrentiev regularization is generated by

$$g_\alpha(t) = t^{-1}\left(1 - (\alpha/(\alpha + 2t))^p\right),$$

and has qualification p.

2.2. General source condition

Note, that it is natural [23] when dealing with ill-posed operator equation $Af = u$, $A = A^* \geq 0$, to assume that its solution is in $Range(\varphi(A))$. This assumption is called as the source condition, or the source condition generated by an index function φ. For example, the results of [8] and [14] correspond to the source condition of Hölder type generated by the index function $\varphi(t) = t^r$, and $A = L_K : \mathcal{H}_K \to \mathcal{H}_K$. The results of [8, 14] have been recently improved and generalized in [24] by analysing the general regularization scheme and more general index functions φ defined as follows.

Definition 3. *A function* $\varphi : [0, d] \to \mathbb{R}$ *is called an operator monotone index function if* $\varphi(0) = 0$*, and for any pair of non-negative self-adjoint operators* A_1, A_2 *such that* $\|A_1\|, \|A_2\| \leq d$ *and* $A_1 \leq A_2$ *one has* $\varphi(A_1) \leq \varphi(A_2)$.

We denote by Ψ_d the class of functions satisfying Definition 3.

We also introduce the class Φ_d of index functions $\varphi : [0, d] \to \mathbb{R}$ admitting a splitting as $\varphi = \varphi_1 \cdot \varphi_2$, where $\varphi_2 \in \Psi_d$ and φ_1 is a non-decreasing Lipschitz function such that $\varphi_1(0) = 0$.

The splitting $\varphi = \varphi_1 \cdot \varphi_2$ is not unique. Therefore, without loss of generality we may assume that the Lipschitz constant for φ_1 is equal to 1 that, in particular, allows the following bound (see, e.g. [11], p. 209):

$$\|\varphi_1(L_K) - \varphi_1(\frac{1}{m^2} S_{\mathbf{x}}^* \mathbb{D} S_{\mathbf{x}})\|_{\mathcal{H}_K \to \mathcal{H}_K} \leq \|L_K - \frac{1}{m^2} S_{\mathbf{x}}^* \mathbb{D} S_{\mathbf{x}}\|_{\mathcal{H}_K \to \mathcal{H}_K}. \quad (9)$$

Below, we present some examples of index functions $\varphi \in \Psi_d \cup \Phi_d$:

- $\varphi(t) = t^r$; $\varphi \in \Psi_d$ for $r \in (0, 1]$, $d > 0$.

- $\varphi(t) = \log^{-r} \frac{1}{t}$; $\varphi \in \Psi_d$ for $r \in (0,1]$, $d \in (0,1)$.
- $\varphi(t) = \log^{-r} \log \frac{1}{t}$; $\varphi \in \Psi_d$ for $r \in (0,1]$, $d \in (0,1)$.
- $\varphi(t) = t^p \log^{-r} \frac{1}{t}$; $\varphi \notin \Psi_d$, but $\varphi \in \Phi_d$ for $p \geq 1$, $r \in [0,1]$, $d \in (0,1)$.

Following [12] we say that the *qualification* p *covers* a function $\varphi \in \Psi_d \cup \Phi_d$ if the function $\frac{t^p}{\varphi(t)}$ is non-decreasing for $t \in [0, d]$.

In the sequel the performance of the regularized ranking will be analyzed under the assumption that $f_\rho \in Range(\varphi(L_K))$, $\varphi \in \Psi_d \cup \Phi_d$, that includes, as a particular case, the source condition $f_\rho \in Range(L_K^r)$, which was used in [8, 13, 14].

2.3. Linear combination of regularized rankers

For a set of regularized solutions $\{f_{\mathbf{z}}^{\alpha_i}\}_{i=1}^I$ we consider a linear combination

$$f_{\mathbf{z}} = \sum_{p \in \Pi} c_p f_{\mathbf{z}}^{\alpha_p}, \quad \Pi \subset \{1, 2, \ldots, I\}.$$

The idea of a combination of rankers is also used in boosting, which is one of the most popular methods for ranking problems [4]. In boosting the so-called weak rankers are trained sequentially and then blended in a linear composition. Although boosting is usually explored with large ensembles, it seems to be less effective if we want to build an ensemble from a small set of already strong rankers. For example, the experiments in [22] show that the regularized least squares ranking algorithm outperforms boosting of threshold functions considered as weak rankers.

Let $\overline{c} = (\overline{c}_p)$ be the vector of the ideal coefficients for the approximation of the target function by the above mentioned linear combination in a Hilbert space $\mathcal{H}$, such that $\overline{c} = (\overline{c}_p)$ solves the minimization problem

$$\|f_\rho - \sum_{p \in \Pi} c_p f_{\mathbf{z}}^{\alpha_p}\|_{\mathcal{H}} \to \min \tag{10}$$

Then the vector $\overline{c} = (\overline{c}_p)$ also solves the linear system $Gc = \overline{g}$ with the Gram matrix $G = \left(\langle f_{\mathbf{z}}^{\alpha_p}, f_{\mathbf{z}}^{\alpha_j} \rangle_{\mathcal{H}}\right)_{p,j \in \Pi}$ and the right-hand-side vector $\overline{g} = \left(\langle f_\rho, f_{\mathbf{z}}^{\alpha_p} \rangle_{\mathcal{H}}\right)_{p \in \Pi}$.

In contrast to G, the vector $\overline{g}$ is not accessible, since it depends on the unknown solution of the ill-posed equation $L_K f = L_K f_\rho$.

At the same time, the regularization theory tells (see, e.g. [11], Proposition 2.15, [21]) that the values of linear bounded functionals $l(\cdot) = \langle l, \cdot \rangle_{\mathcal{H}}$ at the solution of an ill-posed operator equation can be estimated by the so-called linear functional strategy (LFS) much more accurately that the solution in $\mathcal{H}$. In the next section we extend LFS to the ranking context.

3. LFS based on the general regularization scheme

3.1. Preliminaries

At first we present some facts that will be used in our analysis.

Proposition 1. *Qualification 1 covers any index function $\varphi \in \Psi_d$.*

Proof. From [11] we know that $\varphi \in \Psi_d$ admits an integral representation of the form

$$\varphi(t) = at + \int_0^\infty \frac{t}{(t+\lambda)\lambda} \mu(d\lambda),$$

where a is come constant and μ is a finite positive measure that does not have mass on $(0, d)$. Then

$$\frac{t}{\varphi(t)} = \left[a + \int_0^\infty \frac{\mu(d\lambda)}{(t+\lambda)\lambda} \right]^{-1},$$

and it is clear that $\frac{t}{\varphi(t)}$ is a non-decreasing function of t, which means that φ is covered by qualification 1. □

Proposition 2 ([11], Proposition 2.22). *Let $\varphi \in \Psi_d$. Then for each $d' \in (0, d]$ there is a positive number $c = c(d', \varphi)$ such that for any pair of non-negative self-adjoint operators A, B, satisfying the bounds $\|A\|_{X\to X}, \|B\|_{X\to X} \leq d'$ in some Hilbert space X, we have:*

$$\|\varphi(A) - \varphi(B)\|_{X\to X} \leq c\varphi(\|A - B\|_{X\to X}).$$

Proposition 3 ([11], Proposition 2.7). *Let $\{g_\alpha\}$ be a regularization family described in Definitions 1, 2. If an index function $\varphi : [0, d] \to \mathbb{R}$ is covered by the qualification of $\{g_\alpha\}$, then*

$$\sup_{0\leq t\leq d} |1 - g_\alpha(t)t|\varphi(t) \leq \max\{\gamma_0, \gamma_p\}\varphi(\alpha).$$

Proposition 4. *Assume, that an index function $\varphi : [0, d] \to \mathbb{R}$ is covered by a qualification p and admits a splitting $\varphi = \varphi_1 \cdot \varphi_2$, where φ_i is a non-decreasing function and $\varphi_i(0) = 0$, $i = 1, 2$. Then φ_1, φ_2 are also covered by the qualification p.*

Proof. Indeed, for the function φ_2, say, and for any $0 < t_1 < t_2 \leq d$ we have:

$$\frac{t_1^p}{\varphi_2(t_1)} = \frac{\varphi_1(t_1)t_1^p}{\varphi(t_1)} \leq \frac{\varphi_1(t_1)t_2^p}{\varphi(t_2)} \leq \frac{\varphi_1(t_2)t_2^p}{\varphi(t_2)} = \frac{t_2^p}{\varphi_2(t_2)}.$$

□

Proposition 5. *Let $\varphi \in \Psi_d$. Then*

$$\sup_{0\leq t\leq d} |g_\alpha(t)\varphi(t)| \leq (1 + \gamma_0 + \gamma_{-1})\frac{\varphi(\alpha)}{\alpha}.$$

Proof. Observe, that for $\varphi \in \Psi_d$, $t \in [0, d]$, it holds

$$\frac{\varphi(t)}{t + \alpha} \leq \frac{\varphi(\alpha)}{\alpha},$$

because for $t \leq \alpha$ we have

$$\frac{\varphi(t)}{t + \alpha} \leq \frac{\varphi(t)}{\alpha} \leq \frac{\varphi(\alpha)}{\alpha}.$$

At the same time, Proposition 1 tells us that φ is covered by qualification 1, which means that $\frac{\varphi(t)}{t}$ is a non-increasing function, and for $\alpha < t$ it holds

$$\frac{\varphi(t)}{t + \alpha} \leq \frac{t}{t + \alpha} \cdot \frac{\varphi(t)}{t} \leq \frac{\varphi(t)}{t} \leq \frac{\varphi(\alpha)}{\alpha}.$$

Then from Definition 1 it follows that

$$\begin{aligned}\sup_{0\leq t\leq d} |g_\alpha(t)\varphi(t)| &\leq \sup_{0\leq t\leq d} \frac{\varphi(t)}{t + \alpha} \cdot \sup_{0\leq t\leq d} |g_\alpha(t)(t + \alpha)| \\ &\leq \frac{\varphi(\alpha)}{\alpha}\left[\sup_{0\leq t\leq d} |1 - g_\alpha(t)t| + 1 + \sup_{0\leq t\leq d} \alpha|g_\alpha(t)|\right] \\ &\leq (1 + \gamma_0 + \gamma_{-1})\frac{\varphi(\alpha)}{\alpha}.\end{aligned}$$

□

Proposition 6 ([8, 24]). *Assume, that $f_\rho \in \mathcal{H}_K$ and consider $L_{\mathbf{x}} = \frac{1}{m^2} S_{\mathbf{x}}^* \mathbb{D} S_{\mathbf{x}} : \mathcal{H}_K \to \mathcal{H}_K$. Then for any $\delta \in (0,1)$ with confidence $1-\delta$ it holds that*

$$\|L_K - L_{\mathbf{x}}\|_{\mathcal{H}_K \to \mathcal{H}_K} \le \frac{26\kappa^2 c_\delta}{\sqrt{m}},$$
$$\|L_K f_\rho - \frac{1}{m^2} S_{\mathbf{x}}^* \mathbb{D}\mathbf{y}\|_{\mathcal{H}_K} \le \frac{26\kappa M c_\delta}{\sqrt{m}}, \tag{11}$$

where $\kappa = \sup_{x \in X} \sqrt{K(x,x)}$, and $c_\delta = \max\{\log \frac{2}{\delta}, 1\}$.

As a direct consequence of Proposition 6, it also holds that

$$\begin{aligned}\|L_{\mathbf{x}} f_\rho - \frac{1}{m^2} S_{\mathbf{x}}^* \mathbb{D}\mathbf{y}\|_{\mathcal{H}_K} &\le \|(L_K - L_{\mathbf{x}}) f_\rho\|_{\mathcal{H}_K} + \|L_K f_\rho - \frac{1}{m^2} S_{\mathbf{x}}^* \mathbb{D}\mathbf{y}\|_{\mathcal{H}_K} \\ &\le \frac{26\kappa c_\delta}{\sqrt{m}} \left(\kappa \|f_\rho\|_{\mathcal{H}_K} + M\right) = \frac{\mathbf{C}}{\sqrt{m}} \log \frac{1}{\delta}. \end{aligned} \tag{12}$$

Here and in the sequel the generic symbol $\mathbf{C}$ stands for a scalar coefficient that may depend on $\rho, K, M, \gamma_{-1}, \gamma_0, \gamma_p$, but does not depend on α, m, δ. The value of $\mathbf{C}$ may be different at different places.

In the subsequent analysis, we will assume that $f_\rho \in \mathit{Range}(\varphi(L_K))$, $\varphi \in \Phi_d \cup \Psi_d$. Moreover, a function $\Theta(t) = \varphi(t)t$ will play some role. Note, that from the very definition of Θ it follows that

$$\lim_{t \to 0} \frac{\Theta(t)}{t} = \lim_{t \to 0} \frac{t}{\Theta^{-1}(t)} = 0.$$

In particular, for sufficiently large m with confidence $1-\delta$ we have

$$\Theta^{-1}(m^{-\frac{1}{2}}) \ge 26\kappa^2 c_\delta m^{-\frac{1}{2}} \ge \|L_K - L_{\mathbf{x}}\|_{\mathcal{H}_K \to \mathcal{H}_K}. \tag{13}$$

Theorem 7. *Assume that $f_\rho \in \mathit{Range}(\varphi(L_K))$, $\varphi \in \Phi_d \cup \Psi_d$, and $l \in \mathit{Range}(\psi(L_K))$, $\psi \in \Psi_d$, $d > \sup_x |K(x,x)| \left(2 + 26 m^{-\frac{1}{2}} c_\delta\right)$. Assume also that m is large enough such that (13) is satisfied. If the qualification p of a regularization family $\{g_\alpha\}$ covers the index function $\varphi \cdot \psi$, then for $f_{\mathbf{z}}^\alpha = g_\alpha(L_{\mathbf{x}}) \frac{1}{m^2} S_{\mathbf{x}}^* \mathbb{D}\mathbf{y}$ and $\alpha = \Theta^{-1}(m^{-\frac{1}{2}})$ with confidence $1-\delta$ we have*

$$|\langle l, f_\rho \rangle_{\mathcal{H}_K} - \langle l, f_{\mathbf{z}}^\alpha \rangle_{\mathcal{H}_K}| = O\left(\varphi\left(\Theta^{-1}(m^{-1/2})\right) \psi\left(\Theta^{-1}(m^{-1/2})\right) \log \frac{1}{\delta}\right),$$

where the coefficient implicit in O-symbol depends on l, f_ρ, g_α, but does not depend on m and δ.

Proof. We consider the case $\varphi \in \Phi_d$. In the case $\varphi \in \Psi_d$ the argument is the same.

Let $r_\alpha(t) = 1 - g_\alpha(t)t$. Then for $l = \psi(L_K)v$, $v \in \mathcal{H}_K$, we have

$$|\langle l, f_\rho\rangle_{\mathcal{H}_K} - \langle l, f_z^\alpha\rangle_{\mathcal{H}_K}| = I_{\rho,\mathbf{x}} + I_{\rho,\mathbf{z}}, \tag{14}$$

where

$$\begin{aligned} I_{\rho,\mathbf{x}} &= |\langle v, \psi(L_K) r_\alpha(L_\mathbf{x}) f_\rho\rangle_{\mathcal{H}_K}|, \\ I_{\rho,\mathbf{z}} &= |\langle v, \psi(L_K) g_\alpha(L_\mathbf{x})(L_\mathbf{x} f_\rho - m^{-2} S_x^* \mathbb{D}Y)\rangle_{\mathcal{H}_K}|. \end{aligned}$$

By using the assumption that $f_\rho = \varphi(L_K)u$, $u \in \mathcal{H}_K$, we can continue as follows

$$\begin{aligned} I_{\rho,\mathbf{x}} &= |\langle v, \psi(L_K) r_\alpha(L_\mathbf{x}) f_\rho\rangle_{\mathcal{H}_K}| \\ &\le \mathbf{C}\|(\psi(L_K) - \psi(L_\mathbf{x}) + \psi(L_\mathbf{x})) r_\alpha(L_\mathbf{x})(\varphi(L_K) - \varphi(L_\mathbf{x}) + \varphi(L_\mathbf{x}))\|_{\mathcal{H}_K \to \mathcal{H}_K} \\ &\le \mathbf{C}(I^1_{\rho,\mathbf{x}} + I^2_{\rho,\mathbf{x}} + I^3_{\rho,\mathbf{x}} + I^4_{\rho,\mathbf{x}}), \end{aligned} \tag{15}$$

where

$$\begin{aligned} I^1_{\rho,\mathbf{x}} &= \|\psi(L_\mathbf{x}) r_\alpha(L_\mathbf{x}) \varphi(L_\mathbf{x})\|_{\mathcal{H}_K \to \mathcal{H}_K}, \\ I^2_{\rho,\mathbf{x}} &= \|\psi(L_\mathbf{x}) r_\alpha(L_\mathbf{x})(\varphi(L_K) - \varphi(L_\mathbf{x}))\|_{\mathcal{H}_K \to \mathcal{H}_K}, \\ I^3_{\rho,\mathbf{x}} &= \|(\psi(L_K) - \psi(L_\mathbf{x})) r_\alpha(L_\mathbf{x}) \varphi(L_\mathbf{x})\|_{\mathcal{H}_K \to \mathcal{H}_K}, \\ I^4_{\rho,\mathbf{x}} &= \|(\psi(L_K) - \psi(L_\mathbf{x})) r_\alpha(L_\mathbf{x})(\varphi(L_K) - \varphi(L_\mathbf{x}))\|_{\mathcal{H}_K \to \mathcal{H}_K}. \end{aligned}$$

The assumption that $\{g_\alpha\}$ covers the index function $\varphi(\cdot)\psi(\cdot) \in \Psi_d \cup \Phi_d$, together with Proposition 3, gives us that

$$I^1_{\rho,\mathbf{x}} \le \sup_{0 \le t \le d} |r_\alpha(t)\varphi(t)\psi(t)| \le \max\{\gamma_0, \gamma_p\}\varphi(\alpha)\psi(\alpha).$$

To estimate $I^2_{\rho,\mathbf{x}}$ we recall that by the definition of Φ_d the function $\varphi \in \Phi_d$ allows a splitting as $\varphi = \varphi_1 \cdot \varphi_2$, where $\varphi_2 \in \Psi_d$ and φ_1 is such that (9) holds. Then

$$\varphi(L_K) - \varphi(L_\mathbf{x}) = \varphi_1(L_\mathbf{x})(\varphi_2(L_K) - \varphi_2(L_\mathbf{x})) + \varphi_2(L_K)(\varphi_1(L_K) - \varphi_1(L_\mathbf{x})),$$

and

$$I^2_{\rho,\mathbf{x}} \le I^{2,1}_{\rho,\mathbf{x}} + I^{2,2}_{\rho,\mathbf{x}},$$

$$I^{2,1}_{\rho,\mathbf{x}} = \|\psi(L_{\mathbf{x}})r_\alpha(L_{\mathbf{x}})\varphi_1(L_{\mathbf{x}})(\varphi_2(L_K) - \varphi_2(L_{\mathbf{x}}))\|_{\mathcal{H}_K\to\mathcal{H}_K},$$
$$I^{2,2}_{\rho,\mathbf{x}} = \|\psi(L_{\mathbf{x}})r_\alpha(L_{\mathbf{x}})\varphi_2(L_K)(\varphi_1(L_K) - \varphi_1(L_{\mathbf{x}}))\|_{\mathcal{H}_K\to\mathcal{H}_K}.$$

For $\alpha = \Theta^{-1}(m^{-\frac{1}{2}})$ Propositions 2-4, together with (13), give us

$$\begin{aligned} I^{2,1}_{\rho,\mathbf{x}} &\le \|r_\alpha(L_{\mathbf{x}})\varphi_1(L_{\mathbf{x}})\psi(L_{\mathbf{x}})\|_{\mathcal{H}_K\to\mathcal{H}_K}\varphi_2\left(\|L_K - L_{\mathbf{x}}\|_{\mathcal{H}_K\to\mathcal{H}_K}\right) \\ &\le \mathbf{C}\varphi_1(\alpha)\psi(\alpha)\varphi_2(\alpha) = \mathbf{C}\psi(\alpha)\varphi(\alpha). \end{aligned}$$

Moreover, using Propositions 3, 4, 6 and 9 we obtain

$$I^{2,2}_{\rho,\mathbf{x}} \le \mathbf{C}\varphi_2(d)\sup_{0\le t\le d}|r_\alpha(t)\psi(t)|m^{-\frac{1}{2}}c_\delta \le \mathbf{C}\psi(\alpha)m^{-\frac{1}{2}}\log\frac{1}{\delta}.$$

Then

$$I^2_{\rho,\mathbf{x}} \le \mathbf{C}(\varphi(\alpha)\psi(\alpha) + \psi(\alpha)m^{-\frac{1}{2}}\log\frac{1}{\delta}).$$

In the same way, one can show that $I^3_{\rho,\mathbf{x}} \le \mathbf{C}\varphi(\alpha)\psi(\alpha)$, $I^4_{\rho,\mathbf{x}} \le \mathbf{C}(\varphi(\alpha)\psi(\alpha) + \psi(\alpha)m^{-\frac{1}{2}}\log\frac{1}{\delta})$, and in view of (15) it gives us the bound

$$I_{\rho,\mathbf{x}} \le \mathbf{C}(\varphi(\alpha)\psi(\alpha) + \psi(\alpha)m^{-\frac{1}{2}}\log\frac{1}{\delta}). \tag{16}$$

On the other hand,

$$I_{\rho,\mathbf{z}} \le \mathbf{C}\|\psi(L_K)g_\alpha(L_{\mathbf{x}})(L_{\mathbf{x}}f_\rho - \frac{1}{m^{-2}}S^*_x\mathbb{D}Y)\|_{\mathcal{H}_K} \le \mathbf{C}(I^1_{\rho,\mathbf{z}} + I^2_{\rho,\mathbf{z}}),$$

where

$$I^1_{\rho,\mathbf{z}} = \|(\psi(L_K) - \psi(L_{\mathbf{x}}))g_\alpha(L_{\mathbf{x}})(L_{\mathbf{x}}f_\rho - m^{-2}S^*_x\mathbb{D}Y)\|_{\mathcal{H}_K},$$
$$I^2_{\rho,\mathbf{z}} = \|\psi(L_{\mathbf{x}})g_\alpha(L_{\mathbf{x}})(L_{\mathbf{x}}f_\rho - m^{-2}S^*_x\mathbb{D}Y)\|_{\mathcal{H}_K}.$$

Recall that $\psi \in \Psi_d$. Then from Propositions 2, 6, (8), (12) and (13) it follows that

$$I^1_{\rho,\mathbf{z}} \le \mathbf{C}\gamma_{-1}\psi(\|L_K - L_{\mathbf{x}}\|_{\mathcal{H}_K\to\mathcal{H}_K})\frac{m^{-\frac{1}{2}}\log\frac{1}{\delta}}{\alpha} \le \mathbf{C}\frac{\psi(\alpha)}{\alpha}m^{-\frac{1}{2}}\log\frac{1}{\delta}.$$

Moreover, Proposition 5, together with (12), gives us the bound

$$I^2_{\rho,\mathbf{z}} \le \mathbf{C}\sup_{0\le t\le d}|\psi(t)g_\alpha(t)|m^{-\frac{1}{2}}\log\frac{1}{\delta} \le \mathbf{C}\frac{\psi(\alpha)}{\alpha}m^{-\frac{1}{2}}\log\frac{1}{\delta}.$$

Then

$$I_{\rho,\mathbf{z}} \le \mathbf{C}\frac{\psi(\alpha)}{\alpha}m^{-\frac{1}{2}}\log\frac{1}{\delta}. \tag{17}$$

Observe now that for $\alpha = \Theta^{-1}(m^{-\frac{1}{2}})$

$$\varphi(\alpha) = \frac{m^{-\frac{1}{2}}}{\alpha} \ge m^{-\frac{1}{2}}.$$

Therefore, from (14), (16), (17) we finally receive the requested error bound

$$\begin{aligned}|\langle l, f_\rho\rangle_{\mathcal{H}_K} - \langle l, f_z^\alpha\rangle_{\mathcal{H}_K}| &\le \mathbf{C}\psi(\alpha)\left(\varphi(\alpha) + \frac{m^{-\frac{1}{2}}}{\alpha}\right)\log\frac{1}{\delta}\\ &= O(\varphi(\Theta^{-1}(m^{-\frac{1}{2}}))\psi(\Theta^{-1}(m^{-\frac{1}{2}}))\log\frac{1}{\delta}).\end{aligned} \tag{18}$$

□

Remark 2. *Observe that from (14) it follows that in (18) a coefficient implicit in O-symbol may be taken the same for all functionals l allowing the representation in the form $l = \psi(L_K)v$, $\|v\|_{\mathcal{H}_K} \le R$, where R is a fixed constant. This observation will be useful for the estimation of the excess risk.*

Theorem 8. *Assume the conditions on f_ρ and m in Theorem 7 hold. If the qualification of the regularization family covers the index function $\varphi(t)\sqrt{t}$, then for $\alpha = \Theta^{-1}(m^{-1/2})$ with confidence $1-\delta$ we have*

$$\mathcal{E}(f_{\mathbf{z}}^\alpha) - \mathcal{E}(f_\rho) = O\left(\varphi^2\left(\Theta^{-1}(m^{-1/2})\right)\Theta^{-1}(m^{-1/2})\log^2\frac{1}{\delta}\right).$$

Proof. Observe that for any $f \in \mathcal{H}_K$ one can write

$$\begin{aligned}&\int_X\int_X [f(x) - f(x')]^2 d\rho_X(x)d\rho_X(x')\\ &= \int_X\int_X (f(x) - f(x'))\langle K_x - K_{x'}, f\rangle_{\mathcal{H}_K} d\rho_X(x)d\rho_X(x')\\ &= 2\int_X\int_X f(x)\langle K_x - K_{x'}, f\rangle_{\mathcal{H}_K} d\rho_X(x)d\rho_X(x')\\ &= 2\langle L_K f, f\rangle_{\mathcal{H}_K} = 2\|L_K^{\frac{1}{2}}f\|_{\mathcal{H}_K}^2.\end{aligned}$$

Then from (1) it follows that

$$\mathcal{E}(f_{\mathbf{z}}^{\alpha}) - \mathcal{E}(f_\rho) = \int_X \int_X [f_{\mathbf{z}}^{\alpha}(x) - f_{\mathbf{z}}^{\alpha}(x') - f_\rho(x) + f_\rho(x')]^2 d\rho_X(x) d\rho_X(x')$$
$$= 2\|L_K^{\frac{1}{2}}(f_{\mathbf{z}}^{\alpha} - f_\rho)\|_{\mathcal{H}_K}^2. \tag{19}$$

Now using Remark 2 and Theorem 7 with $l \in Range(\psi(L_K))$, $\psi(t) = \sqrt{t}$ we obtain

$$\|L_K^{\frac{1}{2}}(f_{\mathbf{z}}^{\alpha} - f_\rho)\|_{\mathcal{H}_K} = \sup_{\|v\|_{\mathcal{H}_K}=1} |\langle L_K^{\frac{1}{2}} v, f_\rho\rangle_{\mathcal{H}_K} - \langle L_K^{\frac{1}{2}} v, f_{\mathbf{z}}^{\alpha}\rangle_{\mathcal{H}_K}|$$
$$= O\left(\varphi\left(\Theta^{-1}(m^{-1/2})\right)\sqrt{\Theta^{-1}(m^{-1/2})}\log\frac{1}{\delta}\right),$$

and this relation, together with (19), gives us the statement of the theorem. □

Remark 3. *For $\varphi(t) = t^r$ Theorem 8 gives an estimation of the excess risk of order $O\left(m^{-\frac{2r+1}{2r+2}}\right)$ that essentially improves the order $O\left(m^{-\frac{r}{2r+3}}\right)$ given by [8].*

3.2. Ranking by the Linear Functional Strategy in $\mathcal{H}_K$ (RLFS-$\mathcal{H}_K$)

For our further discussion it is instructive to compare Theorem 7 with a learning rate for ranking

$$\|f_\rho - f_{\mathbf{z}}^{\alpha}\|_{\mathcal{H}_K} = O(\varphi(\Theta^{-1}(m^{-\frac{1}{2}}))\log\frac{1}{\delta}) \tag{20}$$

obtained in [24] under the conditions of Theorem 7. This comparison tells us that for $l \in \mathcal{H}_K$ the guaranteed accuracy in estimating the value $\langle f_\rho, l\rangle_{\mathcal{H}_K}$ if of higher order than that for estimating f_ρ in $\mathcal{H}_K$.

Theorem 7 also tells us that the accuracy of order $o\left(\varphi\left(\Theta^{-1}(m^{-1/2})\right)\log\frac{1}{\delta}\right)$ in approximating $\langle f_\rho, f_{\mathbf{z}}^{\alpha_p}\rangle_{\mathcal{H}_K}$ can be achieved with the same value of the regularization parameter α that does not depend on $f_{\mathbf{z}}^{\alpha_p}$.

This observation opens the door for applying one's favorite parameter choice rule, and it may be done only once.

Thus, due to Theorem 7 the vector $\overline{g} = \left(\langle f_\rho, f_{\mathbf{z}}^{\alpha_p}\rangle_{\mathcal{H}_K}\right)_{p\in\Pi}$ can be approximated by a vector $\tilde{g} = \left(\langle f_{\mathbf{z}}^{\alpha}, f_{\mathbf{z}}^{\alpha_p}\rangle_{\mathcal{H}_K}\right)_{p\in\Pi}$ such that

$$\|\overline{g} - \tilde{g}\|_{\mathbb{R}^q} = o\left(\varphi\left(\Theta^{-1}(m^{-1/2})\right)\log\frac{1}{\delta}\right),$$

where the coefficient implicit in o-symbol depends on the cardinality q of the involved sequence of the regularized ranking functions $\{f_{\mathbf{z}}^{\alpha_p}\}_{p\in\Pi}$, which is assumed not to be very large.

This means that the linear function strategy allows us to construct a ranking function

$$f_{\mathbf{z}} = \sum_{p\in\Pi} \tilde{c}_p f_{\mathbf{z}}^{\alpha_p}, \quad \tilde{c} = (\tilde{c}_p) = G^{-1}\tilde{g},$$

such that under the condition of Theorem 7 with confidence $1-\delta$ it holds

$$\|f_\rho - f_{\mathbf{z}}\|_{\mathcal{H}_K} = \min_{c_p} \|f_\rho - \sum_{p\in\Pi} c_p f_{\mathbf{z}}^{\alpha_p}\|_{\mathcal{H}_K} + o\left(\varphi\left(\Theta^{-1}(m^{-1/2})\right)\log\frac{1}{\delta}\right).$$

In other words, we can effectively construct a ranking function from $span\{f_{\mathbf{z}}^{\alpha_p}\}$, whose distance in $\mathcal{H}_K$ to a risk minimizer differs from the minimal one by a quantity of higher order than the guaranteed convergence rate (20).

3.3. Ranking by the Linear Functional Strategy in $L_2(X,\rho_X)$ (RLFS-L_2)

If the minimization problem (10) is considered in the space $\mathcal{H} = L_2(X,\rho_X)$, then neither Gram matrix $G = \left(\langle f_{\mathbf{z}}^{\alpha_p}, f_{\mathbf{z}}^{\alpha_j}\rangle_{\mathcal{H}}\right)_{p,j\in\Pi}$, nor the vector $\overline{g} = \left(\langle f_\rho, f_{\mathbf{z}}^{\alpha_p}\rangle_{\mathcal{H}}\right)_{p\in\Pi}$ is accessible, since the marginal probability distribution ρ_X is not assumed to be given.

This issue can be resolved if we aim at approximating the other risk minimizer

$$\overline{f_\rho} = f_\rho - \int_X f_\rho(x)d\rho_X(x).$$

Proposition 9. *Assume that the conditions of Theorem 7 hold. Then for $\alpha_p, \alpha_j \geq m^{-\frac{1}{2}}$ with confidence $1-\delta$ we have*

$$\langle\overline{f_\rho}, f_{\mathbf{z}}^{\alpha_p}\rangle_{L_2(X,\rho_X)} = m^{-2}\langle\mathbb{D}\mathbf{y}, S_{\mathbf{x}}f_{\mathbf{z}}^{\alpha_p}\rangle_{\mathbb{R}^m} + O(m^{-\frac{1}{2}}\log\tfrac{1}{\delta}),$$
$$\langle f_{\mathbf{z}}^{\alpha_p}, f_{\mathbf{z}}^{\alpha_j}\rangle_{L_2(X,\rho_X)} = m^{-1}\langle S_{\mathbf{x}}f_{\mathbf{z}}^{\alpha_p}, S_{\mathbf{x}}f_{\mathbf{z}}^{\alpha_j}\rangle_{\mathbb{R}^m} + O(m^{-\frac{1}{2}}\log\tfrac{1}{\delta}),$$

where the coefficients implicit in O-symbol do not depend on p, j, m and δ.

Proof. Let $I_{\mathcal{H}_K} : \mathcal{H}_K \to L_2(X, \rho_X)$ be the inclusion operator of $\mathcal{H}_K$ into $L_2(X, \rho_X)$. It is known (see, e.g., [11], Lemma 4.1, and references therein) that with confidence $1 - \delta$ it holds

$$\|I^*_{\mathcal{H}_K} I_{\mathcal{H}_K} - \frac{1}{m} S^*_{\mathbf{x}} S_{\mathbf{x}}\|_{\mathcal{H}_K \to \mathcal{H}_K} \leq \mathbf{C} m^{-\frac{1}{2}} \log \frac{1}{\delta}, \tag{21}$$

where

$$I^*_{\mathcal{H}_K} f(\cdot) = \int_X K(x, \cdot) f(x) d\rho_X(x).$$

Then it is easy to see that

$$\begin{aligned} I^*_{\mathcal{H}_K} \overline{f_\rho}(\cdot) &= \int_X K(x, \cdot) \left(f_\rho(x) - \int_X f_\rho(x') d\rho_X(x') \right) d\rho_X(x) \\ &= \int_X \int_X K(x, \cdot) f_\rho(x) d\rho_X(x) d\rho_X(x') \\ &- \int_X \int_X K(x', \cdot) f_\rho(x) d\rho_X(x) d\rho_X(x') = L_K f_\rho. \end{aligned}$$

Moreover, since $f^{\alpha_p}_{\mathbf{z}} \in \mathcal{H}_K$, we have

$$\begin{aligned} \langle \overline{f_\rho}, f^{\alpha_p}_{\mathbf{z}} \rangle_{L_2(X,\rho_X)} &= \langle \overline{f_\rho}, I_{\mathcal{H}_K} f^{\alpha_p}_{\mathbf{z}} \rangle_{L_2(X,\rho_X)} = \langle I^*_{\mathcal{H}_K} \overline{f_\rho}, f^{\alpha_p}_z \rangle_{\mathcal{H}_K} = \langle L_K f_\rho, f^{\alpha_p}_{\mathbf{z}} \rangle_{\mathcal{H}_K} \\ &= m^{-2} \langle S^*_{\mathbf{x}} \mathbb{D}\mathbf{y}, f^{\alpha_p}_{\mathbf{z}} \rangle_{\mathcal{H}_K} + \langle L_K f_\rho - m^{-2} S^*_{\mathbf{x}} \mathbb{D}\mathbf{y}, f^{\alpha_p}_{\mathbf{z}} \rangle_{\mathcal{H}_K} \\ &= \frac{1}{m^2} \langle \mathbb{D}\mathbf{y}, S_{\mathbf{x}} f^{\alpha_p}_z \rangle_{\mathbb{R}^m} + I_m, \end{aligned} \tag{22}$$

where the term I_m can be estimated with the use of Proposition 6 as follows

$$\begin{aligned} |I_m| &= |\langle L_K f_\rho - m^{-2} S^*_{\mathbf{x}} \mathbb{D}\mathbf{y}, f^{\alpha_p}_{\mathbf{z}} \rangle_{\mathcal{H}_K}| \\ &\leq \|L_K f_\rho - m^{-2} S^*_{\mathbf{x}} \mathbb{D}\mathbf{y}\|_{\mathcal{H}_K} \|f^{\alpha_p}_{\mathbf{z}}\|_{\mathcal{H}_K} \leq \frac{26 \kappa M c_\delta}{\sqrt{m}} \|f^{\alpha_p}_{\mathbf{z}}\|_{\mathcal{H}_K} \end{aligned} \tag{23}$$

Recall that $f^{\alpha}_{\mathbf{z}} = g_\alpha(L_{\mathbf{x}}) \frac{1}{m^2} S^*_{\mathbf{x}} \mathbb{D}\mathbf{y}$. Therefore, for $\alpha > m^{-\frac{1}{2}}$ the bounds (7), (8) and Proposition 6 give us

$$\begin{aligned} \|f^{\alpha}_{\mathbf{z}}\|_{\mathcal{H}_K} &\leq \|g_\alpha(L_{\mathbf{x}}) L_{\mathbf{x}}\|_{\mathcal{H}_K \to \mathcal{H}_K} \|f_\rho\|_{\mathcal{H}_K} \\ &+ \|g_\alpha(L_{\mathbf{x}})\|_{\mathcal{H}_K \to \mathcal{H}_K} \|L_K - L_{\mathbf{x}}\|_{\mathcal{H}_K \to \mathcal{H}_K} \|f_\rho\|_{\mathcal{H}_K} \\ &+ \|g_\alpha(L_{\mathbf{x}})\|_{\mathcal{H}_K \to \mathcal{H}_K} \|L_K f_\rho - m^{-2} S^*_{\mathbf{x}} \mathbb{D}\mathbf{y}\|_{\mathcal{H}_K} \\ &\leq (\gamma_0 + 1) \|f_\rho\|_{\mathcal{H}_K} + \frac{26 \kappa c_\delta \gamma_{-1}}{\alpha \sqrt{m}} (\kappa \|f_\rho\|_{\mathcal{H}_K} + M) \end{aligned}$$

Then the first relation of the proposition follows from (22), (23). The second relation can be proved by similar argument but instead of the bound on $\|L_K - L_{\mathbf{x}}\|$ one needs to use (21). $\square$

Remark 4. *Note, that in view of Theorem 7 one expects α to be of order $\Theta^{-1}(m^{-\frac{1}{2}}) \geq m^{-\frac{1}{2}}$. Therefore, the assumption of Proposition 9 that $\alpha_p, \alpha_j \geq m^{-\frac{1}{2}}$ is not restrictive. The proposition tells us that the unknown quantities $\langle \overline{f_\rho}, f_{\mathbf{z}}^{\alpha_p} \rangle_{L_2(X,\rho_X)}$, $\langle f_{\mathbf{z}}^{\alpha_p}, f_{\mathbf{z}}^{\alpha_j} \rangle_{L_2(X,\rho_X)}$ can be effectively estimated with the best parametric rate $O(m^{-\frac{1}{2}})$ without any additional regularization.*

Let us continue with approximating the solution of the minimization problem (10) in the space $\mathcal{H} = L_2(X, \rho_X)$. Recall that this problem can be reduced to the linear system $Gc = \bar{g}$ with $G = \left(\left\langle f_{\mathbf{z}}^{\alpha_p}, f_{\mathbf{z}}^{\alpha_j} \right\rangle_{L_2(X,\rho_X)} \right)_{p,j \in \Pi}$, $\bar{g} = \left(\left\langle \bar{f}_\rho, f_{\mathbf{z}}^{\alpha_p} \right\rangle_{L_2(X,\rho_X)} \right)_{p \in \Pi}$. As it has been already mentioned, the number q of the elements in the set $\left\{ f_{\mathbf{z}}^{\alpha_p} \right\}_{p \in \Pi}$ is assumed to be negligible compared to the cardinality m of the training set, such that q-dependent coefficients do not affect the orders $o(\varphi(\theta^{-1}(m^{-1/2})))$ or $O(m^{-1/2})$. Some applications, where this is the case are presented in the next section.

In view of Proposition 9 the matrix $\tilde{G} = \left(m^{-1} \left\langle S_{\mathbf{x}} f_{\mathbf{z}}^{\alpha_p}, S_{\mathbf{x}} f_{\mathbf{z}}^{\alpha_j} \right\rangle_{\mathbb{R}^m} \right)_{p,j \in \Pi}$ and the vector $\tilde{g} = \left(m^{-2} \left\langle \mathbb{D}\mathbf{y}, S_{\mathbf{x}} f_{\mathbf{z}}^{\alpha_p} \right\rangle_{\mathbb{R}^m} \right)_{p \in \Pi}$ can be considered as approximations of G and $\bar{g}$ respectively. Then Proposition 9 tells us that with confidence $1-\delta$ it holds

$$\left\| G - \tilde{G} \right\|_{\mathbb{R}^q \to \mathbb{R}^q} = O(m^{-1/2} \log \frac{1}{\delta}), \quad \|\bar{g} - \tilde{g}\|_{\mathbb{R}^q} = O(m^{-1/2} \log \frac{1}{\delta}). \tag{24}$$

With the matrix $\tilde{G}$ in hand one can easily check whether or not $\tilde{G}^{-1}$ exists. If it exists then in view of (24) it is natural to assume that for sufficiently large m with confidence $1 - \delta$ we have

$$\left\| G - \tilde{G} \right\|_{\mathbb{R}^q \to \mathbb{R}^q} < \frac{1}{\left\| \tilde{G}^{-1} \right\|_{\mathbb{R}^q \to \mathbb{R}^q}}. \tag{25}$$

This assumption allows the application of the well-known Banach theorem on inverse operators (see, e.g., [25], V. 4.5), which tells that

$$\left\|G^{-1}\right\|_{\mathbb{R}^q\to\mathbb{R}^q} \le \frac{\left\|\tilde{G}^{-1}\right\|_{\mathbb{R}^q\to\mathbb{R}^q}}{1-\left\|\tilde{G}^{-1}\right\|_{\mathbb{R}^q\to\mathbb{R}^q}\left\|G-\tilde{G}\right\|_{\mathbb{R}^q\to\mathbb{R}^q}} = O(1). \tag{26}$$

Consider the vectors $\bar{c} = G^{-1}\bar{g}, \tilde{c} = \tilde{G}^{-1}\tilde{g}$. Then from (24)–(26) it follows that

$$\|\bar{c}-\tilde{c}\|_{\mathbb{R}^q} = O(m^{-1/2}\log\frac{1}{\delta}). \tag{27}$$

Theorem 10. *Consider a ranking function* $f_{\mathbf{z}} = \sum_{p\in\Pi}\tilde{c}_p f_{\mathbf{z}}^{\alpha_p}, \tilde{c} = (\tilde{c})_{p\in\Pi} = \tilde{G}^{-1}\tilde{g}$*, and assume (25) and the conditions of Proposition 9. Then with confidence* $1-\delta$ *it holds*

$$\left\|\overline{f_\rho} - f_{\mathbf{z}}\right\|_{L_2(X,\rho_X)} \le \min_{c_p}\left\|\overline{f_\rho} - \sum_{p\in\Pi} c_p f_{\mathbf{z}}^{\alpha_p}\right\|_{L_2(X,\rho_X)} + O(m^{-1/2}\log\frac{1}{\delta}),$$

where a coefficient implicit in O-symbol may depend on ρ *and the cardinality of* Π*, but does not depend on* m *and* δ*.*

Proof. It is clear that

$$\left\|\overline{f_\rho} - f_{\mathbf{z}}\right\|_{L_2(X,\rho_X)} \le \min_{c_p}\left\|\overline{f_\rho} - \sum_{p\in\Pi} c_p f_{\mathbf{z}}^{\alpha_p}\right\|_{L_2(X,\rho_X)} + q\left\|\bar{c}-\tilde{c}\right\|_{\mathbb{R}^q}\max_p\left\|f_{\mathbf{z}}^{\alpha_p}\right\|_{L_2(X,\rho_X)} \tag{28}$$

Since $\mathcal{H}_K$ is assumed to be embedded in $L_2(X,\rho_X)$, for any $f\in\mathcal{H}_K$ we have $\|f\|_{L_2(X,\rho_X)} \le c\|f\|_{\mathcal{H}_K}$. Moreover, in the proof of Proposition 9 it has been shown that for $\alpha_p \ge m^{-1/2}$ the norms $\left\|f_{\mathbf{z}}^{\alpha_p}\right\|_{\mathcal{H}_K}$ are uniformly bounded. Then the statement of the theorem follows from (27) and (28). □

In the next section we demonstrate the effectiveness of the ranking functions $f_{\mathbf{z}}$ constructed by means of the linear functional strategy.

4. Numerical experiments

4.1. Academic example

In our first experiment we are going to demonstrate an advantage of the linear functional strategy ranking algorithm compared to a regularized

Solution	Pairwise Misranking	Mean Square Error
$f_{\mathbf{z}}^{\alpha_1}$	6.04%	0.0020
$f_{\mathbf{z}}^{\alpha_2}$	7.58%	0.0032
$f_{\mathbf{z}}^{\alpha_3}$	7.89%	0.0041
$f_{\mathbf{z}}$	1.18%	0.0003

Table 1: Performance of regularized ranking functions $f_{\mathbf{z}}^{\alpha_p}, p = 1, 2, 3$, with the fixed values of the regularization parameters and their linear combination $f_{\mathbf{z}}$ constructed according to RLFS-L_2.

ranking with fixed regularization parameters. As in [26], we consider a target function

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

and construct its regularized approximations $f_{\mathbf{z}}^{\alpha_p}$ in RKHS associated with the kernel $K(x, u) = xu + e^{-8(x-u)^2}$. In this experiment the approximations $f_{\mathbf{z}}^{\alpha_p}$ are constructed by means of Lavrentiev regularization (5) for $\alpha = \alpha_p = 0.1p, p = 1, 2, 3$, so that $q = 3$.

A training set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ is formed by $m = 21$ noisy values $y_i = f_\rho(x_i) + \varepsilon_i$ at points x_i sampled randomly from uniform distribution on $[0, 2\pi]$; noise values ε_i are sampled randomly from uniform distribution on $[-0.02, 0.02]$. A ranking function $f_{\mathbf{z}}$ is constructed from $\left\{f_{\mathbf{z}}^{\alpha_p}\right\}_{p=1}^3$ according to RLFS-L_2 without any additional regularization. Then a test set of 100 samples $\{x_j, f_\rho(x_j)\}_{j=1}^{100}$ is constructed to measure the performance of $f_{\mathbf{z}}^{\alpha_p}, p = 1, 2, 3$, and $f_{\mathbf{z}}$ in terms of the percentage of pairwise misranked points x_j and in terms of the mean square error. The results are presented in Table 1.

Moreover, Figure 1 displays the graphs of the constructed approximations $f_{\mathbf{z}}^{\alpha_p}, p = 1, 2, 3$, $f_{\mathbf{z}}$ and f_ρ. From these table and figure it can be concluded that the ranking function $f_{\mathbf{z}}$ essentially outperforms the approximations $f_{\mathbf{z}}^{\alpha_p}$ from which it has been constructed.

4.2. Experiments with MovieLens 20-40, 40-60, 60-80

The datasets MovieLens are publicly available from the following URL: `http://www.grouplens.org/taxonomy/term/14`. These datasets were previously used in [22] for comparing the performance of ranking algorithms such as the magnitude-preserving ranking, and RankBoost [4].

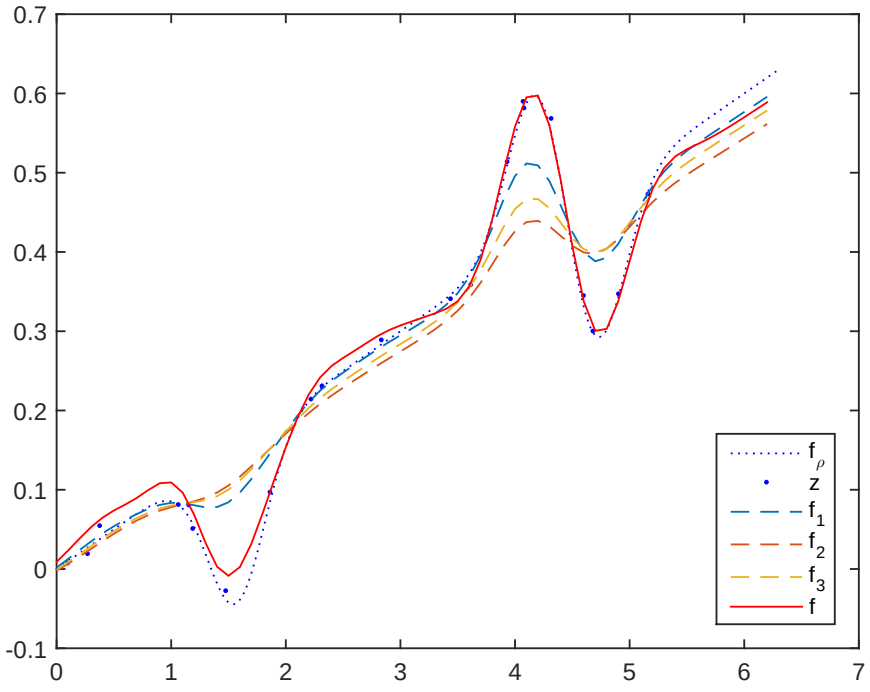

Figure 1: The target function f_ρ (blue dotted line), the training set $\mathbf{z}$, the ranking functions $f_{\mathbf{z}}^{\alpha}$ for $\alpha = 0.1, 0.2, 0.3$, and $f_{\mathbf{z}}$ (red line).

MovieLens dataset consists of anonymous ratings of approximately 3900 movies made by 6040 users. In the experiment 300 users were randomly selected as the test reviewers from those who rated $50 - 300$ movies.

We followed exactly the experimental set-up of [22] and [4]. For each of the test reviewers, a training set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$, $25 \le m \le 150$, contains this reviewer rating y_i of the i-th movie, and a vector $x_i = (x_i^1, x_i^2, \ldots, x_i^{300})$ of ratings of the same movie made by 300 randomly chosen users, which are considered as reference reviewers.

The learning task consists in predicting the rating y made by the considered test reviewer for a movie that was not included in the training set. The prediction input is a vector $x = (x^1, x^2, \ldots, x^{300})$ of ratings made by the reference reviewers for the movie under consideration.

The rating performance has been measured in terms of the fraction of misranked pairs in the set of $m \in [25, 150]$ movies that have not been included in the training set.

The ranking functions $f_{\mathbf{z}}^{\alpha}$ are constructed in $\mathcal{H}_K$, $K(x, \overline{x}) = \exp(-\|x - \overline{x}\|^2_{\mathbb{R}^{300}}/1000)$, by Lavrentiev regularization with $\alpha \in \{\alpha_i = (0.97)^i, i = 0, 1, \ldots, 200\}$. For ranking by the linear functional strategy (RLFS) we use $f_{\mathbf{z}}^{\alpha_p}$, $p = 0, 152, 200$, i.e. $q = 3$, because such α_p have three different orders of magnitude $10^0, 10^{-2}, 10^{-3}$. In our numerical tests the value α for approximating $\langle f_\rho, f_{\mathbf{z}}^{\alpha_p}\rangle_{\mathcal{H}_K}$ by $\langle f_{\mathbf{z}}^{\alpha}, f_{\mathbf{z}}^{\alpha_p}\rangle_{\mathcal{H}_K}$ was selected randomly from $\{\alpha_i = (0.97)^i, i = 0, 1, \ldots, 200\}$. The strategy is performed in two versions:

- in $\mathcal{H}_K$ (RLFS-$\mathcal{H}_K$),
- in $L_2(X, \rho_X)$ (RLFS-L_2),

Performance of RLFS is compared against RankBoost algorithm by [4], and against $f_{\mathbf{z}}^{\alpha}$ chosen from the whole sequence $\{f_{\mathbf{z}}^{\alpha_i}\}_{i=0}^{200}$ by means of the cross-validation with respect to the fraction of misranked pairs (our performance metric).

Table 2 reports the mean values of the performance metric over 10 simulations and 300 different test reviewers. It can be seen from the Table 2 that RLFS-$\mathcal{H}_K$ outperforms the competitors. Moreover, RLFS-L_2 outperforms RankBoost.

4.3. Ranking by linear functional strategy in application to the prediction of Nocturnal Hypoglycemia (NH)

NH (a low blood glucose (BG) concentration during the sleep period) is the most common and particular worrisome hypoglycemia in individuals

Ranking algorithm	Data set		
	MovieLens $20-40$	MovieLens $40-60$	MovieLens $60-80$
RLFS-$\mathcal{H}_K$	0.4034	0.3820	0.3833
RLFS-L_2	0.4251	0.4064	0.3983
$f_{\mathbf{z}}^{\alpha}$ + cross-validation	0.4153	0.3920	0.3933
RankBoost	—	0.4760	0.4631

Table 2: Performance in the terms of fraction of misranked pairs in the test set

with diabetes.

One of the first method for predicting NH was proposed in [27]. The method is based on the latest before bed BG-measurement x (mg/dL), and it ranks the risk of NH by means of a ranking function

$$f_a(x) = \begin{cases} 1, & x < a \text{ (mg/dL)}, \\ -1, & x \geq a \text{ (mg/dL)}, \end{cases}$$

where the value (-1) means no risk of NH, while 1 means that there is a risk of NH.

In clinical study [27] the ranking functions $f_a = f_{a_i}$, $a_i = 90 + 18(i-1)$, $i = 1, 2, \ldots, 6$, were tested on a data set consisting of 71 nights; NH was observed in 34% of them.

As a result, $f_{a_3}, a_3 = 126\ (mg/dL)$ was suggested as the best NH-predictor among $\{f_a\}$.

Assuming that there is an ideal ranking function $f_\rho(x)$ predicting NH from the latest before bed BG-measurement x, the idea is to approximate $\overline{f_\rho}(x)$ by

$$f_{\mathbf{z}} = \sum_{p=1}^{6} \tilde{c}_p f_{a_p}(x),$$

where the coefficients $\tilde{c}_p$ are calculated according to RLFS-L_2 with the use of a training set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ of historical data such that $y_i = 1$, if the latest before bed measurement x_i was followed by a night with NH, and $y_i = -1$, if this was not the case.

Thus, we define the coefficients vector $\tilde{c} = (\tilde{c_p})_{p=1}^6$ as the solution of the linear system $\tilde{G}\tilde{c} = \tilde{g}$, where

$$\tilde{G} = (m^{-1}\langle S_{\mathbf{x}} f_{a_p}, S_{\mathbf{x}} f_{a_q} \rangle_{\mathbb{R}^m})_{p,q=1}^6, \ \tilde{g} = (m^{-2}\langle \mathbb{D}\mathbf{y}, S_{\mathbf{x}} f_{a_p} \rangle_{\mathbb{R}^m})_{p=1}^6.$$

We use a data set collected withing EU-project DIAdvisor (`www.diadvisor.eu`). The set consists of 150 nights; NH was observed in 27% of them. We consider 200 training sets $\mathbf{z}$ that have been randomly chosen from the DIAdvisor data set.

Each training set $\mathbf{z}$ consisting of $m = 70$ nights has been used to construct $f_{\mathbf{z}}(x)$ by means of the above mentioned linear functional strategy. Then the constructed ranking function $f_{\mathbf{z}}(x)$ has been tested on the other 80 nights that were not included in $\mathbf{z}$.

Ranking function $f_{\mathbf{z}}$ was transformed to classification $f_{0,1}$ in the following way:

$$f_{0,1}(x) = \begin{cases} 1, \text{ if } f_{\mathbf{z}}(x) \geq 0 \\ -1, \text{ otherwise.} \end{cases}$$

Following [27] the performance of NH-predictors, such as $f_{\mathbf{z}}(x)$, or $f_{a_p}(x)$, $p = 1, 2, \ldots, 6$ has been evaluated in terms of Sensitivity (SE), Specificity (SP), Positive Predictive Value(PPV), Negative Predictive Value (NPV), and also in terms of F_1 score. The average values of the above performance metrics over all 200 tests are reported in Table 3.

Ranking function	SE (%)	SP(%)	PPV (%)	NPV (%)	F_1
f_{a_1}	49.21	99.1	95.1	84.06	0.6400
f_{a_2}	69.82	91.5	75.36	89.08	**0.7200**
f_{a_3}	79.49	71.32	50.61	90.38	0.6141
f_{a_4}	83.99	53.28	39.87	90.01	0.5370
f_{a_5}	97.26	38.61	36.88	97.43	0.5317
f_{a_6}	97.26	31.09	34.23	96.86	0.5033
$f_{\mathbf{z}}$	71.32	85.92	68.07	89.15	**0.6824**

Table 3: Comparative Performance of NH-predictors

As it can be seen from the table, the ranking function f_{a_3}, that was suggested in [27] as the best NH-predictor, is the fourth worst in our tests. On the other hand, the ranking function f_{a_2}, that was the second worst in the tests [27], is the best in our experiments.

At the same time, the ranking function $f_{\mathbf{z}}$, that has been constructed by means of the linear functional strategy on the basis of all considered ranking functions f_{a_i}, $i = 1, 2, \ldots, 6$ exhibits the second best performance.

This can be seen as a demonstration of the ability of the linear functional strategy to construct a predictor that automatically follows the leader. Such a predictor looks more safe than the individual predictors from which it is constructed.

Acknowledgment

The authors are supported by the Austrian Fonds Zur Forderung der Wissenschaftlichen Forschung (FWF), grant P25424 "Data-driven and Problem-Oriented Choice of the Regularization Space", and by D-A-CH Project I1669 "Multi-parameter Regularization for Lifting the Curse of Dimensionality".

References

[1] W. Cohen, R. Schapire, Y. Singer, Learning to order things, J. Artif. Intell. Res. 10 (1999) 243–270.

[2] R. Herbrich, T. Graepel, K. Obermayer, Large margin rank boundaries for ordinal regression, in: P. J. Bartlett, B. Schlkopf, D. Schuurmans, A. J. Smola (Eds.), Advances in Large Margin Classifiers, MIT Press, 2000, pp. 115–132.

[3] K. Crammer, Y. Singer, Pranking with ranking, in: Advances in Neural Information Processing Systems 14, MIT Press, 2001, pp. 641–647.

[4] Y. Freund, R. Iyer, R. E. Schapire, Y. Singer, An efficient boosting algorithm for combining preferences, J. Mach. Learn. Res. 4 (2003) 933–969.

[5] S. Mukherjee, D.-X. Zhou, Learning coordinate covariances via gradients, J. Machine Learning Res. 7 (2006) 519–549.

[6] D. Cossock, T. Zhang, Subset ranking using regression, in: G. Lugosi, H. Simon (Eds.), Learning Theory, volume 4005 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2006, pp. 605–619.

[7] S. Agarwal, P. Niyogi, Generalization bounds for ranking algorithms via algorithmic stability, J. of Mach. Learn. Res. 10 (2009) 441–474.

[8] H. Chen, The convergence rate of a regularized ranking algorithm, Journal of Approximation Theory 164 (2012) 1513 – 1519.

[9] S. Smale, D.-X. Zhou, Learning theory estimates via integral operators and their approximations, Constructive Approximation 26 (2007) 153–172.

[10] V. Kurkova, Learning from data as an optimization and inverse problem, in: K. Madani, A. Dourado Correia, A. Rosa, J. Filipe (Eds.), Computational Intelligence, volume 399 of *Studies in Computational Intelligence*, Springer Berlin Heidelberg, 2012, pp. 361–372.

[11] S. Lu, S. V. Pereverzev, Regularization Theory for Ill-posed Problems. Selected Topics., De Gruyter, Berlin, Boston, 2013.

[12] P. Mathé, S. V. Pereverzev, Geometry of linear ill-posed problems in variable hilbert scales, Inverse Problems 19 (2003) 789–803.

[13] Y. Ying, D.-X. Zhou, Online Pairwise Learning Algorithms with Kernels, ArXiv e-prints (2015).

[14] M. Xu, Q. Fang, S. Wang, Convergence analysis of an empirical eigenfunction-based ranking algorithm with truncated sparsity, Abstract and Applied Analysis 2014 (2014) 197476.

[15] F. Bauer, S. Pereverzev, L. Rosasco, On regularization algorithms in learning theory, J. Complexity 23 (2007) 52–72.

[16] E. De Vito, S. Pereverzyev, L. Rosasco, Adaptive kernel methods using the balancing principle, Foundations of Computational Mathematics 10 (2010) 455–479.

[17] A. Caponnetto, Y. Yao, Cross-validation based adaptation for regularization operators in learning theory, Analysis and Applications 8 (2010) 161–183.

[18] R. Anderssen, The linear functional strategy for improperly posed problems, in: J. Cannon, U. Hornung (Eds.), Inverse Problems, volume 77 of *International Series of Numerical Mathematics*, Birkhäuser, Basel, 1986, pp. 11–30.

[19] A. Goldenshluger, S. V. Pereverzev, Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations, Probability Theory and Related Fields 118 (2000) 169–186.

[20] P. Mathé, S. V. Pereverzev, Direct estimation of linear functionals from indirect noisy observations, Journal of Complexity 18 (2002) 500–516.

[21] F. Bauer, P. Math, S. Pereverzev, Local solutions to inverse problems in geodesy, Journal of Geodesy 81 (2007) 39–51.

[22] C. Cortes, M. Mohri, A. Rastogi, Magnitude-preserving ranking algorithms, in: Proc. of the 24th international conference on Machine learning, pp. 169–176.

[23] P. Mathé, B. Hofmann, How general are general source conditions?, Inverse Problems 24 (2008) 015009.

[24] G. Kriukova, P. Tkachenko, S. Pereverzyev, On the convergence rate and some applications of regularized ranking algorithms, RICAM Reports 2014-21 (2014).

[25] L. Kantorovich, G. Akilov, Functional Analysis, Elsevier Science, 2014.

[26] C. A. Micchelli, M. Pontil, Learning the kernel function via regularization, Journal of Machine Learning Research 6 (2005) 1099–1125.

[27] G. Whincup, R. Milner, Prediction and management of nocturnal hypoglycemia in diabetes, Arch Dis Childhood 62 (1987) 333–337.