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

On the convergence rate and some applications of regularized ranking algorithms

G. Kriukova, S. Pereverzyev, P. Tkachenko

RICAM-Report 2014-21

On the convergence rate and some applications of regularized ranking algorithms

Galyna Kriukova*, Sergei Pereverzyev, Pavlo Tkachenko

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

Abstract

This paper studies the ranking problem in the context of the regularization theory that allows a simultaneous analysis of a wide class of ranking algorithms. Some of them were previously studied separately. For such ones, our analysis gives a better convergence rate compared to the reported in the literature. We also supplement our theoretical results with numerical illustrations and discuss the application of ranking to the problem of estimating the risk from errors in blood glucose measurements of diabetic patients.

Keywords: Ranking, convergence rate, source condition, blood glucose error grid

1. Introduction

In recent years, the ranking problem attracts much attention in the literature [1, 2, 3, 4, 5] because of its importance for the development of new decision making (or recommender) systems. Various applications of ranking algorithms include document retrieval, credit-risk screening, collaborative filtering, recommender systems in electronic commerce and internet applications. However, the ranking problem appears also outside of internet-based technologies. In particu-

*Corresponding author
Email addresses: galyna.kriukova@oeaw.ac.at (Galyna Kriukova), sergei.pereverzyev@oeaw.ac.at (Sergei Pereverzyev), pavlo.tkachenko@oeaw.ac.at (Pavlo Tkachenko)

lar, in diabetes treatment the errors occurring during blood glucose monitoring (BGM) have different risks for patient's health. The problem of estimating the risks from meter errors can be seen as a ranking problem. We will describe this example in more details in Section 4.3.

The ranking problem can be understood as a learning task to compare two different observations and decide which of them is better in some sense. Different types of ranking algorithms are designed to be best suited for the fields of their application, thus for construction of ranking models different approaches are used.

In this paper we consider supervised global ranking function reconstruction. We estimate the quality of a ranking function by its expected ranking error corresponding to the least squares ranking loss. The ranking problem in this setting has been well studied in [1, 2, 6, 7], where a regularization technique in a Reproducing Kernel Hilbert Space (RKHS) has been employed to overcome the intrinsic ill-posedness of this learning problem.

It is well-known that the regularization theory can be profitably used in the context of learning. There is a substantial literature on the use of Tikhonov-Phillips regularization in RKHS for the purpose of supervised learning. Here we refer to [8, 9, 10] and to references therein. A large class of supervised learning algorithms, which are essentially all the linear regularization schemes, has been analyzed in [11].

The starting point of all these investigations is a representation of the supervised learning regression problem as a discretized version of some ill-posed equation in RKHS. Then a regularization scheme is applied to the corresponding normal equation with a self-adjoint positive operator.

In our view, a special feature of the ranking problem differing it from supervised learning regression is that no normalization is required to reduce this problem to an equation with self-adjoint positive operator. As a result, simplified regularization schemes such as Lavrentiev regularization or methods of singular perturbations [12] can be employed to treat the ill-posedness of the ranking problem.

Lavrentiev regularization in RKHS has been analyzed in the context of ranking in [2, 6], while in [7] a ranking based on Spectral cut-off regularization in RKHS has been studied. Moreover, in [6, 7] the convergence rates of the corresponding ranking algorithms have been estimated under the assumption that the ideal target ranking function meets the so-called source condition of Hölder type. It turns out that up to now, in contrast to the situation in supervised learning regression problem [11], only particular regularization methods, such as Lavrentiev or Spectral cut-off, have been employed for ranking, and, moreover, they have been analyzed separately.

In the present study we extend the unified approach of [11] to the ranking problem and estimate the convergence rates of algorithms based on the so-called general simplified regularization scheme. Our analysis not only covers the cases studied in literature [2, 6, 7], but also improves the estimations of the convergence rates [6, 7]. Moreover, the improved estimations are obtained under much more general source conditions.

The paper is organized as follows. In the next section we discuss the problem setting and previous results. In Section 3 we describe the analyzed class of ranking algorithms and estimate their convergence rate. Finally, in the last section we present some numerical illustrations and discuss the application of ranking to the problem of estimating the risks form errors in blood glucose measurements.

2. Problem Setting and Previous Work

Let X be a compact metric space and $Y = [0, M]$, for some $M > 0$. An input $x \in X$ is related to a rank $y \in Y$ through an unknown probability distribution $\rho(x, y) = \rho(y|x)\rho_X(x)$ on $Z = X \times Y$, where $\rho(y|x)$ is the conditional probability of y given x and $\rho_X(x)$ is the marginal probability of x. The distribution ρ is given only through a set of samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$. The ranking problem aims at learning a function $f_{\mathbf{z}} : X \to \mathbb{R}$ that assigns to each input $x \in X$ a rank $f_{\mathbf{z}}(x)$. Then a loss function $l = l\big(f, (x, y), (x', y')\big)$ is utilized to evaluate

the performance of a ranking function $f = f_{\mathbf{z}}$. For given true ranks y and y' of the inputs x, x' the value of l is interpreted as the penalty or loss of f in its ranking of $x, x' \in X$. If x is to be ranking higher than x', such that $y > y'$, and $f(x) > f(x')$, then the loss $l = l\big(f, (x, y), (x', y')\big)$ should be small. Otherwise, the loss will be large.

We further require that l is symmetric with respect to (x, y), (x', y'), and define the risk

$$\mathcal{E}_l(f) = \mathbb{E}_{(x,y),(x',y')\sim\rho}[l\big(f, (x, y), (x', y')\big)]$$

of a ranking function f as the expected value of the loss l with respect to the distribution ρ.

The learning task can be seen as a minimization of the risk, where the choice of the loss function implies the choice of a ranking model. Obviously, the most natural loss function is the following one:

$$l_{0\text{-}1} = 1_{\{(y-y')(f(x)-f(x'))\leq 0\}},$$

or its modification

$$l_m = 1_{\{(y-y')(f(x)-f(x'))<0\}} + \frac{1}{2} \cdot 1_{\{f(x)=f(x')\}}.$$

The empirical 0-1-risk then simply counts the fraction of misranked pairs in the set $\mathbf{z}$ of size m:

$$\widehat{\mathcal{E}}_{0\text{-}1}(f, \mathbf{z}) = \frac{\sum_{i,j=1}^{m} 1_{\{y_i>y_j \wedge f(x_i)\leq f(x_j)\}}}{\sum_{i,j=1}^{m'} 1_{\{y_i>y_j\}}} = \frac{\sum_{i,j:y_i>y_j} 1_{\{f(x_i)-f(x_j)\leq 0\}}}{|\{i, j : y_i > y_j\}|}. \quad (1)$$

However, both l_m and $l_{0\text{-}1}$ loss functions are discontinuous, so the minimization of the empirical risk can be very challenging. As an alternative, in the literature [1, 2, 6, 7] one focuses on the magnitude-preserving least squares loss:

$$l_{mp}^2 = \Big(y - y' - (f(x) - f(x'))\Big)^2,$$

and measures the quality of a ranking function f via the expected risk

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

Note, that $\mathcal{E}(f)$ is a convex functional of f, however a minimizer of (2) is not unique. In the space $L_2(X, \rho_X)$ of square integrable functions with respect to the marginal probability measure ρ_X the risk $\mathcal{E}(f)$ is minimized by a family of functions $f_\rho(x) + c$, where

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

is the so-called target function and c is a generic constant which may take different values at different occurrences. The function $f_\rho(x)$ is also called regression function. Note that $|f_\rho| \leq M$.

However, the target function $f_\rho(x)$ can not be found in practice, because the conditional probability $\rho(y|x)$ is unknown. Therefore, it is convenient to look for a function f from some hypothesis space $\mathcal{H}$ minimizing the approximation error $\|f - f_\rho\|_{\mathcal{H}}$.

A natural choice of a hypothesis space $\mathcal{H} \subset L_2(X, \rho_X)$ is a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H} = \mathcal{H}_K$, which is a Hilbert space of functions $f : X \to \mathbb{R}$ with the property that for each $x \in X$ and $f \in \mathcal{H}_K$ the evaluation functional $e_x(f) := f(x)$ is continuous (i.e. bounded) in the topology of $\mathcal{H}_K$.

It is known (see, e.g., [10, 13]) that every RKHS is generated by a unique symmetric and positive definite continuous function $K : X \times X \to \mathbb{R}$, called the reproducing kernel of $\mathcal{H}_K$, or Mercer kernel. The RKHS $\mathcal{H}_K$ is defined to be a closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ defined as $\langle K_x, K_{\bar{x}} \rangle_K = K(x, \bar{x})$. The reproducing property takes the form $f(x) = \langle f, K_x \rangle_K$.

Let us define $\kappa = \sup_{x \in X} \sqrt{K(x, x)}$. Then $|f| \leq \kappa \|f\|_{\mathcal{H}_K}$.

The RKHS-setting has been used in [2, 6] to define a ranking function

$$f_{\mathbf{z}}^\lambda = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m^2} \sum_{i,j=1}^{m} \big(y_i - y_j - (f(x_i) - f(x_j))\big)^2 + 2\lambda \|f\|_{\mathcal{H}_K}^2 \right\}, \tag{3}$$

and its data-free analogue

$$f^\lambda = \arg\min_{f\in\mathcal{H}_K}\{\mathcal{E}(f) + 2\lambda\|f\|^2_{\mathcal{H}_K}\}.$$

Following [6, 7] we consider the integral operator $L : \mathcal{H}_K \to \mathcal{H}_K$ as

$$Lf = \int_X\int_X f(x)(K_x - K_{\bar{x}})d\rho_X(x)d\rho_X(\bar{x}).$$

It is known (see, e.g. [6]) that the operator L is self-adjoint and positive linear operator on $\mathcal{H}_K$.

In [6] it has been observed that for $f_\rho \in \mathcal{H}_K$ the minimizer f^λ can be written in the following form:

$$f^\lambda = (L + \lambda I)^{-1} Lf_\rho.$$

The latter one can be seen as a Lavrentiev regularized approximation to a solution of the equation

$$Lf = Lf_\rho.$$

On the other hand, it has also been proven in [6] that the minimizer of (3) admits the representation

$$f^\lambda_{\mathbf{z}} = \left(\frac{1}{m^2}S^*_{\mathbf{x}}DS_{\mathbf{x}} + \lambda I\right)^{-1}\frac{1}{m^2}S^*_{\mathbf{x}}D\mathbf{y}, \tag{4}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ and $S_{\mathbf{x}} : \mathcal{H}_K \to \mathbb{R}^m$ is the so-called sampling operator, i.e.

$$S_{\mathbf{x}}(f) = (f(x_1), f(x_2), \ldots, f(x_m))^T,$$

and its adjoint $S^*_{\mathbf{x}} : \mathbb{R}^m \to \mathcal{H}_K$ can be written as

$$S^*_{\mathbf{x}}c = \sum_{i=1}^m c_i K_{x_i},\ c = (c_1, \ldots, c_m)^T,$$

$D = m\mathbf{I} - \mathbf{1}\cdot\mathbf{1}^T$, $\mathbf{y} = (y_1, \ldots, y_m)^T$, and $\mathbf{I}, \mathbf{1}$ are the m-th order unit matrix and the m-th order column vector of all ones respectively.

The same approach was used in [2] and the corresponding discrete approximation was obtained (the approximations are identical up to a transformation

and notation). In [2] the authors also compare and contrast the algorithms for ranking and for supervised learning regression. It is interesting to note that in ranking, as well as in supervised learning regression, one aims at the reconstruction of the same target function f_ρ from a training set $\mathbf{z}$. In [2] the magnitude-preserving ranking (3) is compared with RankBoost (an algorithm designed to minimize the pairwise misranking error [5]), and with kernel ridge regression, which is one of the most studied in supervised learning. The experiment setup is the same as the one described in Section 4. The results show that magnitude-preserving algorithm (3) has benefits over regression and RankBoost algorithms. This comparison leads to an interesting conclusion: although these algorithms have the same unknown target $f_\rho(x)$, their convergence rates to $f_\rho(x)$ may vary.

It is necessary to mention that the convergence rate of a constructed approximation can only be estimated under some a priori assumption on the target f_ρ. In the regularization theory such a priori assumption is usually given in the form of the so-called source condition written in terms of the underlying operator, such as L. For example, in ([6, 7]) it has been assumed that

$$f_\rho \in W_{r,R} := \{f \in \mathcal{H}_K : f = L^r u, \|u\|_{\mathcal{H}_K} \leq R\}.$$

Under such assumption the convergence rate of the algorithm (3) has been estimated in [6] as $O\big(m^{-\frac{r}{2r+3}}\big)$. The same order of the convergence rate and under the same assumption can be derived from [7] for a ranking algorithm based on the spectral cut-off regularization. In the next section we show that in the situations analyzed in [6, 7] the convergence rate can be estimated as $O\big(m^{-\frac{r}{2r+2}}\big)$. This estimation will follow from much more general statement.

3. Ranking algorithms based on the general regularization scheme

A general form of one-parameter regularization algorithms for solving the ranking problem can be defined as follows

$$f_{\mathbf{z}}^{\lambda} = g_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \frac{1}{m^2} S_{\mathbf{x}}^* D\mathbf{y},$$

where $\{g_\lambda\}$ is a one-parameter regularization family.

Note that $f_{\mathbf{z}}^\lambda$ can be seen as the result of the application of the simplified regularization generated by the family $\{g_\lambda\}$ to the discretized version $\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}} f = \frac{1}{m^2} S_{\mathbf{x}}^* D\mathbf{y}$ of the underlying equation $Lf = Lf_\rho$, where the latter is discretized with the use of the training set $\mathbf{z}$.

It is clear that by taking $g_\lambda(t) = (t+\lambda)^{-1}$ we obtain $f_{\mathbf{z}}^\lambda$ defined by (4). Note also that the ranking function $f_{\mathbf{z}}^\lambda$ corresponding to

$$g_\lambda(t) = \begin{cases} \frac{1}{t}, & t \geq \lambda, \\ 0, & 0 \leq t < \lambda. \end{cases}$$

has been studied in [7] and is the result of the regularization by means of spectral cut-off scheme.

Recall, (see, e.g. [14], Definition 2.2) that, in general, a family $\{g_\lambda\}$ is called a regularization on $[0, a]$, if there are constants $\gamma_{-1}, \gamma_{-1/2}, \gamma_0$ for which

$$\begin{aligned} \sup_{0<t\leq a} |1 - t g_\lambda(t)| &\leq \gamma_0, \\ \sup_{0<t\leq a} |g_\lambda(t)| &\leq \frac{\gamma_{-1}}{\lambda}, \\ \sup_{0<t\leq a} \sqrt{t}\, |g_\lambda(t)| &\leq \frac{\gamma_{-1/2}}{\sqrt{\lambda}}. \end{aligned}$$

The maximal p for which

$$\sup_{0<t\leq a} t^p\, |1 - \lambda g_\lambda(t)| \leq \gamma_p \lambda^p$$

is called a qualification of the regularization method generated by a family $\{g_\lambda\}$. Following [15] we also say that the qualification p covers a non-decreasing function $\phi, \phi(0) = 0$, if the function $t \to \frac{t^p}{\phi(t)}$ is non-decreasing for $t \in (0, a]$.

We consider general source conditions of the form

$$f_\rho \in W_{\phi,R} := \left\{ f \in \mathcal{H}_K : f = \phi(L)\, u,\ \|u\|_{\mathcal{H}_K} \leq R \right\},$$

where ϕ is a non-decreasing function such that $\phi(0) = 0$. The function ϕ is called an index function. It is clear that the source condition set $W_{r,R}$ discussed in [6, 7] is a particular case of $W_{\phi,R}$ with $\phi(t) = t^r$.

Note that in general the smoothness expressed through source condition is not stable with respect to perturbations in the involved operator L. As it was mentioned above, only the discrete version $\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}$ of the operator L is available and it is desirable to control $\phi(L) - \phi(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}})$. To meet this desire we follow [16] and consider source condition sets $W_{\phi,R}$ with operator monotone index functions ϕ.

Recall that a function ϕ is operator monotone on $[0, a]$ if for any pair of self-adjoint operators B_1, B_2, with spectra in $[0, a]$ such that $B_1 \leq B_2$ one has $\phi(B_1) \leq \phi(B_2)$. The partial ordering $B_1 \leq B_2$ for self-adjoint operators B_1, B_2 on some Hilbert space $\mathcal{H}$ means that $\forall h \in \mathcal{H}\ \langle B_1 h, h\rangle \leq \langle B_2 h, h\rangle$.

For operator monotone index functions we have the following fact.

Proposition 1 ([14], Proposition 2.21). *Let $\phi : [0, a] \to \mathbb{R}^+$ be operator monotone with $\phi(0) = 0$. For each $0 < a' < a$ there is a constant $c_\phi = c(a', \phi)$ such that for any pair of non-negative self-adjoint operators B_1, B_2 with $\|B_1\|, \|B_2\| \leq a'$ it holds: $\|\phi(B_1) - \phi(B_2)\| \leq c_\phi \phi \|B_1 - B_2\|$.*

This proposition implies that an operator monotone index function can not tend to zero faster, than linearly. For a better convergence rate ϕ may be assumed to be split into a product $\phi(\cdot) = \vartheta(\cdot)\psi(\cdot)$ of a function

$$\psi \in \mathcal{F}_C^a = \left\{\psi : [0, a] \to \mathbb{R}^+, \text{operator monotone}, \psi(0) = 0, \psi(a) \leq C\right\},$$

and monotone Lipschitz function $\vartheta : \mathbb{R}^+ \to \mathbb{R}^+$, $\vartheta(0) = 0$.

The splitting $\phi(\cdot) = \vartheta(\cdot)\psi(\cdot)$ is not unique, therefore we assume that the Lipschitz constant for ϑ is equal to 1 that allows the following bound (see, e.g. [14], p. 209)

$$\left\|\vartheta(L) - \vartheta\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\right\| \leq \left\|L - \frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right\|. \tag{5}$$

It is easy to see that if ϕ is covered by the qualification p of $\{g_\lambda\}$, then ψ, ϑ as well.

The following proposition has been proved in [14].

Proposition 2 ([14], Proposition 2.7). *Let ϕ be any index function and let $\{g_\lambda\}$ be a regularization family of qualification p that covers ϕ. Then*

$$\sup_{0<t\leq a} |1 - tg_\lambda(t)|\, \phi(t) \leq \max\{\gamma_0, \gamma_p\}\phi(\lambda), \quad \lambda \in (0, a].$$

To continue the convergence analysis we introduce the following lemma, proved in [17]:

Lemma 3. *Assume that a space Ξ is equipped with a probability measure μ. Consider $\xi = (\xi_1, \xi_2, \ldots, \xi_m) \in \Xi^m$, where ξ_l, $l = 1, 2, \ldots, m$, are independent random variables, which are identically distributed according to μ. Consider also a map F from Ξ^m into a Hilbert space with the norm $\|\cdot\|$. Assume that F is measurable with respect to a product measure on Ξ^m. If there is $\Delta \geq 0$ such that $\|F(\xi) - \mathbb{E}_{\xi_l} F(\xi)\| \leq \Delta$ for each $1 \leq l \leq m$ and almost every $\xi \in \Xi^m$, then for every $\epsilon > 0$,*

$$Prob_{\xi\in\Xi^m} \left\{\|F(\xi) - \mathbb{E}_\xi(F(\xi))\| \geq \epsilon\right\} \leq 2e^{-\frac{\epsilon^2}{2(\Delta\epsilon + \Sigma^2)}},$$

where $\Sigma^2 = \sum_{l=1}^m \sup_{\xi\setminus\{\xi_l\}\in\Xi^{m-1}} \mathbb{E}_{\xi_l} \left\{\|F(\xi) - \mathbb{E}_{\xi_l} F(\xi)\|^2\right\}$. Moreover, for any $0 < \delta < 1$, with confidence $1 - \delta$ it holds

$$\|F(\xi) - \mathbb{E}_\xi(F(\xi))\| \leq 2(\Delta + \sqrt{\Sigma^2}) \log \frac{2}{\delta}.$$

As it was mentioned above, one needs to control $\left\|\phi\,(L) - \phi(\frac{1}{m^2} S^*_{\mathbf{x}} D S_{\mathbf{x}})\right\|$ through the value of $\phi(t)$ at $t = \|L - \frac{1}{m^2} S^*_{\mathbf{x}} D S_{\mathbf{x}}\|$. For this purpose we prove the following statement

Lemma 4. *For any $0 < \delta < 1$, with confidence $1 - \delta$, it holds that*

$$\left\|L - \frac{1}{m^2} S^*_{\mathbf{x}} D S_{\mathbf{x}}\right\|_{\mathcal{H}_K \to \mathcal{H}_K} \leq \frac{26\kappa^2}{\sqrt{m}} c_\delta,$$

where $c_\delta = \max\left\{\log \frac{2}{\delta}, 1\right\}$.

Proof. In the proof we will use the notation $\|\cdot\|$ for the operator norm $\|\cdot\|_{\mathcal{H}_K \to \mathcal{H}_K}$ for compactness of the expressions. We start with the following bound

$$\begin{aligned}\left\|L-\frac{1}{m^2}S_{\mathbf{x}}^*DS_{\mathbf{x}}\right\| &\leq \left\|\frac{m-1}{m}L-\frac{1}{m^2}S_{\mathbf{x}}^*DS_{\mathbf{x}}\right\|+\left\|\frac{m-1}{m}L-L\right\| \\ &\leq \left\|\frac{m-1}{m}L-\frac{1}{m^2}S_{\mathbf{x}}^*DS_{\mathbf{x}}\right\|+\frac{1}{m}\|L\|\end{aligned}$$

Keeping in mind that $\|K_x\|_{\mathcal{H}_K}^2=\langle K_x(\cdot),K_x(\cdot)\rangle_K=K(x,x)\leq\kappa^2$ it is clear that

$$\|L\|=\max_{f\in\mathcal{H}_K,\|f\|_{\mathcal{H}_K}=1}\left\|\int_X\int_X f(x)(K_x-K_{\overline{x}})d\rho_X(x)d\rho_X(\overline{x})\right\|_{\mathcal{H}_K}\leq 2\kappa^2$$

We continue with the observation from [6] that $\frac{m-1}{m}L=\frac{1}{m^2}\mathbb{E}_{\mathbf{x}}S_{\mathbf{x}}^*DS_{\mathbf{x}}$. Then

$$\left\|L-\frac{1}{m^2}S_{\mathbf{x}}^*DS_{\mathbf{x}}\right\| \leq \left\|\frac{1}{m^2}\mathbb{E}_{\mathbf{x}}S_{\mathbf{x}}^*DS_{\mathbf{x}}-\frac{1}{m^2}S_{\mathbf{x}}^*DS_{\mathbf{x}}\right\|+\frac{2\kappa^2}{m} \tag{6}$$

To estimate the right-hand side of (6) we are going to use Lemma 3 with $\xi=\mathbf{x}$ and $F(\mathbf{x})=\frac{1}{m^2}S_{\mathbf{x}}^*DS_{\mathbf{x}}$. Therefore, for each $1\leq l\leq m$ we consider the following estimation

$$\left\|\frac{1}{m^2}S_{\mathbf{x}}^*DS_{\mathbf{x}}-\frac{1}{m^2}\mathbb{E}_{x_l}S_{\mathbf{x}}^*DS_{\mathbf{x}}\right\|=\max_{\substack{f\in\mathcal{H}_K,\\ \|f\|_{\mathcal{H}_K}=1}}\left\|\frac{1}{m^2}S_{\mathbf{x}}^*DS_{\mathbf{x}}f-\frac{1}{m^2}\mathbb{E}_{x_l}S_{\mathbf{x}}^*DS_{\mathbf{x}}f\right\|_{\mathcal{H}_K}$$

$$=\max_{\substack{f\in\mathcal{H}_K,\\ \|f\|_{\mathcal{H}_K}=1}}\frac{1}{m^2}\left\|\sum_{i=1}^m\sum_{j=1}^m f(x_i)(K_{x_i}-K_{x_j})-\sum_{i=1}^m\sum_{j=1}^m\mathbb{E}_{x_l}f(x_i)(K_{x_i}-K_{x_j})\right\|_{\mathcal{H}_K}.$$

For every $l=1,\ldots,m$ it holds that

$$\mathbb{E}_{x_l}f(x_i)(K_{x_i}-K_{x_j})=\begin{cases}f(x_i)(K_{x_i}-K_{x_j}), & \text{if } i,j\neq l,\\ \mathbb{E}_{x_l}\{f(x_l)K_{x_l}\}-K_{x_j}\mathbb{E}_{x_l}f(x_l), & \text{if } i=l,j\neq l,\\ f(x_i)K_{x_i}-f(x_i)\mathbb{E}_{x_l}K_{x_l}, & \text{if } i\neq l,j=l.\end{cases}$$

Using this we make the following transformation:

$$
\begin{aligned}
& \sum_{i=1}^{m}\sum_{j=1}^{m} f(x_i)(K_{x_i}-K_{x_j}) - \sum_{i=1}^{m}\sum_{j=1}^{m} \mathbb{E}_{x_l} f(x_i)(K_{x_i}-K_{x_j}) \\
= \;& (m-1)f(x_l)K_{x_l} - (m-1)\mathbb{E}_{x_l}\{f(x_l)K_{x_l}\} \\
+ \;& \sum_{i=1, i\neq l}^{m} \big[-f(x_l)K_{x_i} + f(x_i)\big(\mathbb{E}_{x_l}K_{x_l} - K_{x_l}\big) + \mathbb{E}_{x_l} f(x_l)K_{x_i}\big]
\end{aligned}
$$

Note, that each term in the last expression can be bounded by κ^2. For example,

$$
\sup_{\substack{f\in\mathcal{H}_K,\\ \|f\|_{\mathcal{H}_K}\leq 1}} \|\mathbb{E}_{x_l}\{f(x_l)K_{x_l}\}\|_{\mathcal{H}_K} \;=\; \sup_{\substack{f\in\mathcal{H}_K,\\ \|f\|_{\mathcal{H}_K}\leq 1}} \left\| \int_X f(x)K_x d\rho_X(x)\right\|_{\mathcal{H}_K} \leq \kappa^2.
$$

Combining everything together we arrive at the bound

$$
\left\| \frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}} - \frac{1}{m^2}\mathbb{E}_{x_l} S_{\mathbf{x}}^* D S_{\mathbf{x}} \right\| \leq \frac{6(m-1)\kappa^2}{m^2} < \frac{6\kappa^2}{m}
$$

Using the assumption of Lemma 3 that $\|F(\xi) - \mathbb{E}_{\xi_i}F(\xi)\| \leq \Delta$ we obtain an obvious bound

$$
\Sigma^2 = \sum_{i=1}^{m} \sup_{\xi\setminus\{\xi_i\}\in\Xi^{m-1}} \mathbb{E}_{\xi_i}\left\{ \|F(\xi) - \mathbb{E}_{\xi_i}F(\xi)\|^2 \right\} \leq m\Delta^2.
$$

Now applying this lemma to the case when $\xi = \mathbf{x}$, $F(\xi) = \frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}$, $\Delta = \frac{6\kappa^2}{m}$, and $\Sigma^2 \leq m\Delta^2$, we conclude that with confidence $1-\delta$

$$
\begin{aligned}
\left\| \frac{m-1}{m} L - \frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}} \right\| &= \left\| \frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}} - \frac{1}{m^2}\mathbb{E}_{\mathbf{x}} S_{\mathbf{x}}^* D S_{\mathbf{x}} \right\| \\
&\leq 2\Big(\frac{6\kappa^2}{m} + \sqrt{m}\frac{6\kappa^2}{m}\Big)\log\frac{2}{\delta} \leq \frac{24\kappa^2}{\sqrt{m}}\log\frac{2}{\delta} \qquad (7)
\end{aligned}
$$

Substituting the inequality (7) in (6) we prove the required bound. □

Now we are ready to prove the main result of this section.

Theorem 5. *Let $f_\rho \in W_{\phi,R}$, where $\phi(\cdot) = \vartheta(\cdot)\psi(\cdot), \psi \in \mathcal{F}_C^a$, $a > 2\kappa^2\big(1 + 13c_\delta m^{-\frac{1}{2}}\big)$, and ϑ is a monotone function with Lipschitz constant 1, $\vartheta(0) = 0$. Assume also that the regularization family $\{g_\lambda\}$ has a qualification p which covers $\phi(t)$, $t \in [0, a]$. If*

$$
\eta_1 \leq \lambda \leq 1, \qquad (8)
$$

where $\eta_1 := \frac{26\kappa^2}{\sqrt{m}} c_\delta$*, then with confidence* $1-\delta$ *it holds*

$$\left\| f_\rho - f_{\mathbf{z}}^\lambda \right\|_{\mathcal{H}_K} \leq C_1 \phi(\lambda) + C_2 \frac{1}{\lambda\sqrt{m}}, \tag{9}$$

where $C_1 = (1+c_\psi)\max\{\gamma_0, \gamma_p\}R$, $C_2 = 26\kappa^2 c_\delta(\gamma_0 CR + \gamma_{-1}\|f_\rho\|_{\mathcal{H}_K}) + 24\kappa c_\delta M$.

Proof. In the proof we will use the notation $\|\cdot\|$ for the operator norm $\|\cdot\|_{\mathcal{H}_K \to \mathcal{H}_K}$ for compactness of the expressions.

Let

$$r_\lambda(t) = 1 - t g_\lambda(t).$$

We start with the following error decomposition

$$\begin{aligned} f_\rho - f_{\mathbf{z}}^\lambda &= f_\rho - g_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \frac{1}{m^2} S_{\mathbf{x}}^* D\mathbf{y} \\ &= \left(f_\rho - g_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}} f_\rho\right) \\ &+ \left(g_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}} f_\rho - g_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \frac{1}{m^2} S_{\mathbf{x}}^* D\mathbf{y}\right) \end{aligned} \tag{10}$$

Using the assumption that $f_\rho \in W_{\phi,R}$, $\phi(\cdot) = \psi(\cdot)\vartheta(\cdot)$ and the definition of r_λ we can decompose the first term further

$$\begin{aligned} & f_\rho - g_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}} f_\rho = r_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) f_\rho \\ = \;& r_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \phi(L)\, u = r_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \psi(L)\, \vartheta(L)\, u \\ = \;& r_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \phi\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) u \\ + \;& r_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \vartheta\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \left(\psi(L) - \psi\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\right) u \\ + \;& r_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \left(\vartheta(L) - \vartheta\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\right) \psi(L)\, u. \end{aligned}$$

From Proposition 2 we have

$$\begin{aligned} & \left\| r_\lambda\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) \phi\left(\frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}}\right) u \right\|_{\mathcal{H}_K} \\ \leq \;& \max\{\gamma_0, \gamma_p\}\phi(\lambda)\|u\|_{\mathcal{H}_K} \leq \max\{\gamma_0, \gamma_p\}\phi(\lambda) R \end{aligned}$$

Moreover, Proposition 1 allows the bound

$$\begin{aligned}
&\left\| r_\lambda\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\vartheta\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\left(\psi(L)-\psi\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\right)u\right\|_{\mathcal{H}_K}\\
\leq\ &\left\| r_\lambda\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\vartheta\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\right\|_{\mathcal{H}_K} c_\psi\psi\left(\left\|L-\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right\|\right)\|u\|_{\mathcal{H}_K}\\
\leq\ &\max\{\gamma_0,\gamma_p\}\vartheta(\lambda)c_\psi\psi\left(\left\|L-\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right\|\right)R
\end{aligned}$$

Similarly, with the use of (5) we obtain

$$\begin{aligned}
&\left\| r_\lambda\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\left(\vartheta(L)-\vartheta\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\right)\psi(L)\,u\right\|_{\mathcal{H}_K}\\
&\leq \left\| r_\lambda\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\right\|_{\mathcal{H}_K}\left\|L-\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right\|\|\psi(L)\|\|u\|_{\mathcal{H}_K}\\
&\leq \gamma_0 CR\left\|L-\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right\|
\end{aligned}$$

Summing up the above bounds and using Lemma 3 we conclude that for $\lambda \geq \eta_1$ with confidence $1-\delta$ the following holds:

$$\begin{aligned}
&\left\| f_\rho - g_\lambda\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}} f_\rho\right\|_{\mathcal{H}_K}\\
\leq\ &\max\{\gamma_0,\gamma_p\}R\phi(\lambda)+\max\{\gamma_0,\gamma_p\}c_\psi R\vartheta(\lambda)\psi\left(\left\|L-\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right\|\right)\\
+\ &\gamma_0 CR\left\|L-\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right\|,\\
\leq\ &(1+c_\psi)\max\{\gamma_0,\gamma_p\}R\phi(\lambda)+\gamma_0 CR\eta_1. \qquad (11)
\end{aligned}$$

The second term of the decomposition (10) can be bounded as

$$\begin{aligned}
&\left\| g_\lambda\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}} f_\rho - g_\lambda\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\frac{1}{m^2}S_{\mathbf{x}}^* D\mathbf{y}\right\|_{\mathcal{H}_K}\\
\leq\ &\frac{\gamma_{-1}}{\lambda}\left\|\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}} f_\rho - \frac{1}{m^2}S_{\mathbf{x}}^* D\mathbf{y}\right\|_{\mathcal{H}_K}\\
\leq\ &\frac{\gamma_{-1}}{\lambda}\left(\left\|\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}} f_\rho - \frac{m-1}{m}Lf_\rho\right\|_{\mathcal{H}_K}+\left\|\frac{1}{m^2}S_{\mathbf{x}}^* D\mathbf{y}-\frac{m-1}{m}Lf_\rho\right\|_{\mathcal{H}_K}\right).
\end{aligned}$$

From (7) we have

$$\left\| \frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}} f_\rho - \frac{m-1}{m} L f_\rho \right\|_{\mathcal{H}_K} \leq \left\| \frac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}} - \frac{m-1}{m} L \right\| \|f_\rho\|_{\mathcal{H}_K} \leq \eta_1 \|f_\rho\|_{\mathcal{H}_K}$$

To estimate the second summand we will use Lemma 3 with $\xi = \mathbf{z}$ and $F(\xi) = F(\mathbf{z}) = \frac{1}{m^2} S_{\mathbf{z}}^* D\mathbf{y} - \frac{m-1}{m} L f_\rho$. From [6] we know that $\mathbb{E}_{\mathbf{z}} \left(\frac{1}{m^2} S_{\mathbf{x}}^* D\mathbf{y} \right) = \frac{m-1}{m} L f_\rho$. Then by reasoning similar to the proof of Lemma 4 we obtain

$$\|F(\mathbf{z}) - \mathbb{E}_{z_i} F(\mathbf{z})\|_{\mathcal{H}_K} = \left\| \frac{1}{m^2} S_{\mathbf{x}}^* D\mathbf{y} - \mathbb{E}_{z_i} \frac{1}{m^2} S_{\mathbf{x}}^* D\mathbf{y} \right\|_{\mathcal{H}_K} < \frac{6M\kappa}{m}.$$

Now applying Lemma 3 to the case when $\Delta = \frac{6M\kappa}{m}$ and $\Sigma^2 \leq m\Delta^2$, we obtain that with confidence $1 - \delta$

$$\|F(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} F(\mathbf{z})\|_{\mathcal{H}_K} \leq 12\kappa M \left(\frac{1}{m} + \frac{1}{\sqrt{m}} \right) \log \frac{2}{\delta},$$

that is the same as

$$\left\| \frac{1}{m^2} S_{\mathbf{x}}^* D\mathbf{y} - \frac{m-1}{m} L f_\rho \right\|_{\mathcal{H}_K} \leq 12\kappa M \left(\frac{1}{m} + \frac{1}{\sqrt{m}} \right) \log \frac{2}{\delta},$$

because by definition $\mathbb{E}_{\mathbf{z}} F(\mathbf{z}) = 0$. This inequality allows the following bound for the second term of (10)

$$\begin{aligned} &\left\| g_\lambda \left(\tfrac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}} \right) \tfrac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}} f_\rho - g_\lambda \left(\tfrac{1}{m^2} S_{\mathbf{x}}^* D S_{\mathbf{x}} \right) \tfrac{1}{m^2} S_{\mathbf{x}}^* D\mathbf{y} \right\|_{\mathcal{H}_K} \\ &\qquad \leq \tfrac{\gamma_{-1}}{\lambda} (\|f_\rho\|_{\mathcal{H}_K} \eta_1 + \tfrac{24\kappa M c_\delta}{\sqrt{m}}), \end{aligned}$$

which holds with confidence $1 - \delta$. Combining this with (11) we obtain the required estimation

$$\begin{aligned} \left\| f_\rho - f_{\mathbf{z}}^\lambda \right\|_{\mathcal{H}_K} \leq \quad & (1 + c_\psi) \max\{\gamma_0, \gamma_p\} R\phi(\lambda) + \gamma_0 C R \eta_1 \\ & + \frac{\gamma_{-1}}{\lambda} (\|f_\rho\|_{\mathcal{H}_K} \eta_1 + M \frac{24\kappa M c_\delta}{\sqrt{m}}). \end{aligned}$$

□

Remark 1. *Note that the condition similar to (8) has been considered in [9]. This condition just indicates the values of the regularization parameter λ for*

which the error estimate (9) is non-trivial. For example, if $\lambda < \eta_1$, then the right-hand side of (9) becomes larger than a fixed constant, which is not reasonable. Therefore, the condition $\lambda_1 \geq \eta_1$ is not restrictive at all. As to the condition $\lambda \leq 1$, it only simplifies the results and can be replaced by $\lambda \leq a$ for some positive constant a that would eventually appear in the bound.

From Theorem 5 we can immediately derive a data independent (a priori) parameter choice $\lambda_m = \lambda(m)$ and the corresponding convergence rate.

Corollary 6. *Let $\Theta(\lambda) = \phi(\lambda)\lambda$ and*

$$\lambda_m = \Theta^{-1}(m^{-1/2}).$$

Then for sufficiently large $m \in \mathbb{N}$ such that

$$\Theta^{-1}(m^{-1/2})m^{1/2} \geq 26\kappa^2 c_\delta$$

under the assumptions of Theorem 5 with confidence $1-\delta$ we have the following bound

$$\left\|f_\rho - f_{\mathbf{z}}^{\lambda_m}\right\|_{\mathcal{H}_K} \leq (C_1 + C_2)\phi(\Theta^{-1}(m^{-1/2})). \tag{12}$$

Proof. The choice $\lambda = \lambda_m$ balances the two terms in (9), and gives us the required bound. □

Remark 2. *As we already mentioned, the case $f_\rho \in W_{\phi,R}$ with $\phi(t) = t^r$ has been studied in [6, 7]. In this case Corollary 6 guarantees a convergence rate of order $O\left(m^{-\frac{r}{2r+2}}\right)$ for $\lambda_m = m^{-\frac{1}{2r+2}}$. This improves the results [6, 7] where a convergence rate of order $O\left(m^{-\frac{r}{2r+3}}\right)$ has been established for $f_\rho \in W_{r,R}$ and $f_{\mathbf{z}}^\lambda = g_\lambda\left(\frac{1}{m^2}S_{\mathbf{x}}^* D S_{\mathbf{x}}\right)\frac{1}{m^2}S_{\mathbf{x}}^* D\mathbf{y}$ with $g_\lambda(t) = (\lambda + t)^{-1}$ and $g_\lambda(t) = \frac{1}{t}\cdot 1_{\{t\geq\lambda\}}$ respectively.*

4. Numerical illustrations

4.1. Academic example

In our first experiments we are going to show the advantages of the ranking algorithm (3), (4) compared to the supervised learning regression (SLR) where

the same input data and the target function f_ρ appear (see, for example, [9, 11, 14]). Note that in supervised learning regression problem the operator L appearing in the underlying equation has the form

$$Lf = \int_X f(x) K_x d\rho_X(x).$$

Let $x \in X$ be natural numbers from 0 to 100. In our academic example we assume that the rank of each x can be defined as $y = [x/10]$, where the function $[\cdot]$ takes the integer part of its argument $x/10$. As a hypothesis space $\mathcal{H}_K$ we used the RKHS generated by the universal Gaussian kernel [18] $K(x, \bar{x}) = \exp(-\frac{(x-\bar{x})^2}{\gamma})$ with $\gamma = 100$.

The training set was formed by m randomly chosen natural numbers $\{x_i\}_{i=1}^m \subset \{1, 2, \ldots, 100\}$. Such random choice was repeated 10 times for $m = 12, 20, 28$. For each random simulation the training set was separated into two subsets of $m/2$ elements. The first subset was used for constructing the functions $f_{\mathbf{z}}^\lambda$ using the ranking algorithm (3), (4), and the regularized regression learning algorithm [9]. The second subset was then used for adjusting the regularization parameter λ, which was taken from the geometric sequence of 200 numbers $\lambda = \lambda_j = \lambda_0 q^j$ with $\lambda_0 = 1, q = 0.95$. The regularization parameter of our choice minimizes the value of the quantity $\widehat{\mathcal{E}}_{0\text{-}1}$ defined by (1) on the second of the above mentioned subsets.

The constructed functions $f_{\mathbf{z}}^\lambda$ and the corresponding regularization parameters were then taken to test the performance of each method on the set of 100 random inputs. Table 1 reports the result of the comparison: the mean value of the corresponding pairwise misrankings (1) and its standard deviation over 10 simulations.

4.2. MovieLens and Jester Joke Datasets

The datasets MovieLens and Jester Joke are publicly available from the following URL: `http://www.grouplens.org/taxonomy/term/14`. These datasets were previously used for comparing ranking algorithms in [2], where the magnitude-preserving ranking (3), (4) was compared with RankBoost [5], and with SLR.

	Pairwise Misranking			
	Algorithm (4) mean	Algorithm (4) deviation	SLR mean	SLR deviation
m=12	8.16 %	8.04%	16.77%	5.51%
m=20	4.37 %	3.65%	6.84%	5.45%
m=28	1.34 %	1.54%	2.57%	2.62%

Table 1: Comparison of the ranking algorithm (4) with the supervised learning regression algorithm (SLR).

In this subsection we use the above-mentioned datasets to test the performance of one of the ranking algorithms analyzed in Section 3.

Consider

$$f^{\lambda}_{\mathbf{z}} = g_{\lambda}\Big(\frac{1}{m^2} S^*_{\mathbf{x}} D S_{\mathbf{x}}\Big) \frac{1}{m^2} S^*_{\mathbf{x}} D\mathbf{y}, \text{ where } g_{\lambda}(t) = \frac{t+2\lambda}{(\lambda+t)^2}, \tag{13}$$

that corresponds to two times iterated Lavrentiev regularization scheme. To the best of our knowledge, the method (13) has not been discussed yet in the context of ranking, and it is interesting to test it against some of known benchmarks, such as [2].

Recall, that the MovieLens dataset contains 1000209 anonymous ratings of approximately 3900 movies made by 6040 users who visited MovieLens web site (`http://movielens.org`) in the year 2000. Ratings were made on a 5-star scale (whole-star ratings only). Jester Joke dataset contains over 4.1 million continuous anonymous ratings (-10.00 to +10.00) of 100 jokes from 73,421 users.

We followed exactly the experimental set-up of [2], which corresponds to set-up of [5]. For each user, a different predictive model is derived. The ratings of that user are considered as the output values y_i. The other users' ratings of the i-th movie form the i-th input vector x_i. The only difference compared to [2] is that missing movie review values in the input features were not populated with median review score of the given reference reviewer as in [2], but just with -1; and in Jester Joke we shifted all ratings by 10 (so that rating values be non-

negative), and "-1" corresponds to "not rated". This difference only facilitates the computing, because there is no need for data preprocessing by calculating the median scores.

Test reviewers were selected among users who had reviewed between 50 and 300 movies. For a given test reviewer, 300 reference reviewers were chosen at random from one of the three groups and their rating were used to form the input vectors. The groups consist of reviwers who rated $20-40$, $40-60$ and $60-80$ movies/jokes correspondingly. Training was carried out on half of the test reviewer's movie/joke ratings and testing was performed on the other half. The training set was split into 2 halfs: one for constructing the function $f_{\mathbf{z}}^{\lambda}$, another for adjusting the regularization parameter λ from geometric sequence of 200 numbers $\lambda = \lambda_j = \lambda_0 q^j$ with $\lambda_0 = 15$, $q = 0.95$. Gaussian kernel $K(x,\overline{x}) = \exp(-\|x-\overline{x}\|^2_{\mathbb{R}^{300}}/\gamma)$ with $\gamma = 10000$ was chosen. Note that Gaussian kernel $K(x,\overline{x})$ was also used in [2] for constructing $f_{\mathbf{z}}^{\lambda}$ of the form (4) and for SLR, but the value γ was not indicated. We expect that the performance of the ranking algorithms reported in [2] was obtained for optimized values of γ.

The experiment was done for 300 different test reviewers and the average performance was recorded. The whole process was then repeated ten times with different sets of 300 reviewers selected at random. Table 2 reports mean values and standard deviations of pairwise misrankings (1) over these ten repeated experiments for each of the three groups and for each of the tested ranking algorithms. As it can be seen from the table, the ranking algorithm based on the iterated Lavrentiev regularization outperforms the benchmarks.

4.3. Application to blood glucose error grid analysis

The most widely used metric for quantification of the clinical accuracy of blood glucose meters is the Clarke Error Grid Analysis (EGA) developed in 1987 [19]. Since then the researches are trying to improve the Clarke's EGA imposing additional features from clinical practice. The most recent error grid (see Figure 1), called Surveillance Error Grid (SEG), has been introduced in the year 2014 [20]. Within SEG a particular risk is coded by a corresponding

		Pairwise Misranking (1)		
Training inputs	Algorithm (13)	Algorithm (3) as in [2]	SLR [2]	RankBoost [2]
MovieLens 20-40	41.75% ±0.5%	—	—	—
MovieLens 40-60	$39.8\% \pm 0.5\%$	$47.1\% \pm 0.5\%$	$51.1\% \pm 1.1\%$	$47.6\% \pm 0.7\%$
MovieLens 60-80	$38.5\% \pm 0.5\%$	$44.2\% \pm 0.5\%$	$48.4\% \pm 1.3\%$	$46.3\% \pm 1.1\%$
Jester 20-40	$39.3\% \pm 0.7\%$	$41.0\% \pm 0.6\%$	$42.9\% \pm 0.7\%$	$47.9\% \pm 0.8\%$
Jester 40-60	$36.7\% \pm 0.7\%$	$40.8\% \pm 0.6\%$	$42.0\% \pm 0.6\%$	$43.2\% \pm 0.5\%$
Jester 60-80	$35.5\% \pm 0.6\%$	$37.1\% \pm 0.6\%$	$38.5\% \pm 0.6\%$	$41.7\% \pm 0.8\%$

Table 2: Performance of the algorithm (13) and the algorithms tested in [2] in terms of the percentage of pairwise misranking.

color, from green (risk rating = 0) to brown (risk rating = 4). The authors proposed to subdivide the SEG diagram into 8 risk zones corresponding to risk increments of 0.5, and the zones are labeled from "no risk" to "extreme risk" accordingly. In [20] it was mentioned that to build the SEG the authors collected

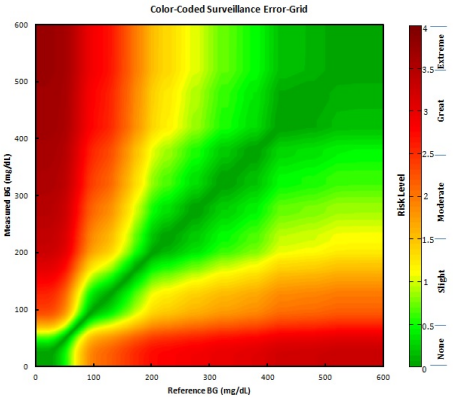


Figure 1: Surveillance Error Grid

the opinions of 234 respondents, among them 206 diabetes clinicians, who rated various treatment scenarios. As a result, 8420 risk ratings were obtained, but among them there were 543 (approximately 6.6%) outliers, so the authors had

to perform a data cleaning procedure to remove inconsistent ratings.

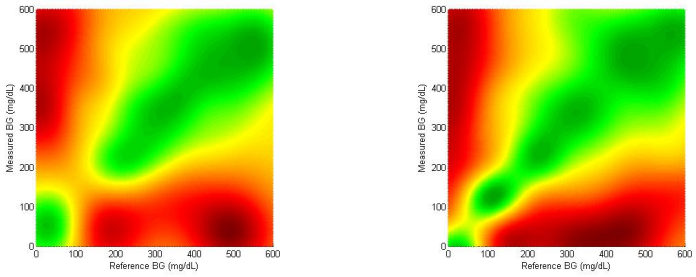
In this subsection we show that within Lavrentiev regularization based ranking one can use only hundreds of ratings instead of thousands to construct an error grid that is almost identical to the SEG. Another potential benefit of the regularized ranking is that it may reduce the outliers effect.

Following [20] we assume that each pair (g_i^r, g_i), where g_i^r denotes a reference value of the blood glucose (BG), and g_i is its corresponding estimate, is related to a risk for patient's health. This risk is considered as a rank of a pair (g_i^r, g_i). The highest risk has a value from a brown region (see Figure 1), and the most safest is the dark green region.

In the experiments reported below we have used the training sets $\mathbf{z}$ containing $m = 100, 200, 300, 400$ random inputs $x_i = (g_i^r, g_i)$, $i = 1, 2, \ldots, m$, uniformly distributed on $[0, 600] \times [0, 600]$, and the corresponding outputs y_i that are the risks assigned to x_i according to SEG.

The ranking functions $f_{\mathbf{z}}^{\lambda}$ have been constructed in the same way as above according to (3), where $\mathcal{H}_K$ is generated by the Gaussian kernel $K(x, \overline{x}) = \exp(-\|x - \overline{x}\|_{\mathbb{R}^2}^2/\gamma)$ with $\gamma = 10000$.

Figure 2 displays BG error grids constructed according to rating functions $f_{\mathbf{z}}^{\lambda}$ trained on training sets with cardinality $m = 100, 200, 300, 400$. As can be seen by comparing this figure with Figure 1, the BG error grid corresponding to the rating function that was trained on the set of 400 risk assessments looks very similar to SEG constructed in [20] with the use of 8240 assessments. Moreover, from Table 3 ($m = 400$) it follows that the assessment according to the BG error grid displayed in Figure 2d may give only 2.9% of the pairwise misranking as compared to SEG, but the majority of these misspecifications corresponds to a rating difference of less than 0.5. This means that in terms of the above mentioned 8 risk zones the assessments according to SEG and the BG error grid from Figure 2d will be similar.



(a) $m = 100$ (b) $m = 200$

(c) $m = 300$ (d) $m = 400$

Figure 2: Reconstruction of SEG using $m = 100, 200, 300, 400$ ranks: $m/2$ as a training set, and $m/2$ for an anjustment of λ

4.4. Regularization parameter choice

It is known that any regularization scheme should be equipped with a strategy for choosing the corresponding regularization parameter. In the above tests the parameters have been chosen on the base of the splitting of the given training sets into two parts. The first parts have been used for constructing the ranking functions $f_{\mathbf{z}}^{\lambda}$, $\lambda = \lambda_j$, $j = 1, 2, \ldots$, while the second parts have been reserved for testing the performance of $f_{\mathbf{z}}^{\lambda_j}$. Then we have chosen $\lambda_+ \in \{\lambda_j\}$ that corresponds to the ranking function $f_{\mathbf{z}}^{\lambda_+}$ exhibiting the best performance on the reserved subsets among the family $\{f_{\mathbf{z}}^{\lambda_j}\}$. Of course, other parameter choice strategies, such as quasi-optimality criterion [21], for example, can be also used in the context of regularized ranking, but for large cardinality m of the training sets such parameter choice strategies may be computationally expensive.

Percentage of cases with rating difference Δy				
Δy	$m = 100$	$m = 200$	$m = 300$	$m = 400$
$0 - 0.5$	76.6 %	88.7 %	95.2 %	96.1 %
$0.5 - 1$	15.6 %	9.2 %	4.0 %	3.0 %
$1 - 1.5$	2.8 %	0.9 %	0.5 %	0.6 %
$1.5 - 2$	1.8 %	0.7 %	0.3 %	0.3 %
$2 - 2.5$	1.6 %	0.3 %	-	-
$2.5 - 3$	0.7 %	0.2 %	-	-
$3 - 3.5$	0.8 %	-	-	-
$3.5 - 4$	0.1 %	-	-	-
Pairwise Misranking	6.74 %	4.0 %	3.0 %	2.9 %

Table 3: Performance of the Lavrentiev regularization based ranking in application to BG error grid analysis

At the same time, Corollary 6 suggests a data independent (a priori) parameter chice $\lambda = \lambda_m = \Theta^{-1}\left(m^{-1/2}\right)$ that balances the two terms in the error bound (9). Of course, this choice requires a knowledge of an index function ϕ describing a source condition $f_\rho \in W_{\phi,R}$, but the latter one does not depend on m, and one can try to approximate it with the use of a training set of small cardinality.

For example, in view of Remark 2, one can try to approximate $\lambda_m = \Theta^{-1}(m^{-1/2})$ by a monomical $\widetilde{\lambda}_m = \alpha m^{-\beta/2}$, $\beta = \frac{1}{r+1}$, where the parameters α, β can be estimated by fitting the function $\widetilde{\lambda}(m) = \alpha m^{-\beta/2}$ to the values of $\lambda_+ = \lambda_+(m)$ that have been found on the base of the splitting of training sets of small cardinality m in the way described above. Then the regularization parameter choice $\lambda = \widetilde{\lambda}(m) = \alpha m^{-\beta/2}$ with the estimated values of α, β can be easily implemented in ranking with an extended training set of larger cardinality m.

We illustrate this approach by the following experiment with the data that have been used in the previous subsection.

We take training subsets with $m = 20, 30, 40, 50$ elements and find corresponding $\lambda_+ = \lambda_+(m)$. Then we consider $\log \lambda_+(m)$, $\log \widetilde{\lambda}(m) = \log \alpha - \frac{1}{2}\beta \log m$, and estimate α, β by solving the system $\log \widetilde{\lambda}(m) = \log \lambda_+(m)$, $m = 20, 30, 40, 50$ for $\log \alpha$ and β in the least squares sense.

The estimated parameters α, β are used to calculate $\lambda = \widetilde{\lambda}(m) = \alpha m^{-\beta/2}$ for $m = 400$. Then the ranking function $f_{\mathbf{z}}^{\lambda}$ has been constructed in the same way as in the previous subsection for $\lambda = \widetilde{\lambda}(400)$ and for the training set $\mathbf{z}$ containing 400 elements.

This experiment has been repeated 20 times and it turns out that the mean value of the pairwise misranking produced by $f_{\mathbf{z}}^{\widetilde{\lambda}(400)}$ on a set of 1000 new, unseen inputs, is 4.27% that is comparable with 2.9% reported in Table 3 for ranking functions $f_{\mathbf{z}}^{\lambda_+(400)}$.

On the other hand, it is clear that the choice $\lambda = \lambda_+(400)$ is computationally much more involved than a priori choice $\lambda = \widetilde{\lambda}(400)$.

The presented experiment demonstrates how a priori regularization parameter choice given by Corollary 6 can be used to reduce the complexity of regularized ranking algorithms.

Acknowledgment

The authors are supported by the Austrian Fonds Zur Forderung der Wissenschaftlichen Forschung (FWF), grant P25424.

References

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

[2] C. Cortes, M. Mohri, A. Rastogi, Magnitude-preserving ranking algorithms, in: Proc. of the 24th international conference on Machine learning, 2007, pp. 169–176.

[3] D. Cossock, T. Zhang, Subset ranking using regression, in: G. Lugosi, H. Simon (Eds.), Learning Theory, Vol. 4005 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2006, pp. 605–619. doi:10.1007/11776420_44.

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

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

[6] H. Chen, The convergence rate of a regularized ranking algorithm, J. of Approx. Theory 164 (12) (2012) 1513–1519. doi:10.1016/j.jat.2012.09.001.

[7] 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. doi:10.1155/2014/197476.

[8] F. Cucker, D. X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Vol. 24 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 2007.

[9] S. Smale, D.-X. Zhou, Learning theory estimates via integral operators and their approximations, Constructive Approx. 26 (2) (2007) 153–172. doi:10.1007/s00365-006-0659-y.

[10] I. Steinwart, A. Christmann, Support Vector Machines, Information Science and Statistics, Springer, New York, 2008.

[11] F. Bauer, S. V. Pereverzyev, L. Rosasco, On regularization algorithms in learning theory, J. of Complex. 23 (1) (2007) 52–72.

[12] F. Liu, M. Z. Nashed, Convergence of regularized solutions of nonlinear ill-posed problems with monotone operators, Vol. 177 of Partial differential equations and applications. Lecture Notes in Pure and Appl. Math., Dekker, New York, 1996.

[13] G. Wahba, Spline Models for Observational Data, Vol. 59 of CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990. doi:10.1137/1.9781611970128.

[14] S. Lu, S. V. Pereverzev, Regularization theory. Selected topics, Vol. 58 of Inverse and Ill-Posed Problems Series, Walter de Gruyter GmbH, Berlin/Boston, 2013.

[15] P. Mathé, S. V. Pereverzev, Geometry of linear ill-posed problems in variable hilbert scales, Inverse Problems 19 (789) (2003) 789803. doi:10.1088/0266-5611/19/3/319.

[16] P. Mathé, S. V. Pereverzev, Moduli of continuity for operator valued functions, Inverse Problems 23 (5-6) (2002) 623–631. doi:10.1081/NFA-120014755.

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

[18] A. Caponnetto, C. A. Michelli, M. Pontil, Y. Ying, Universal multi-task kernels, J. Mach. Learn. Res. 9 (2008) 1615–1646.

[19] W. Clarke, D. Cox, L. Gonder-Frederick, W. Carter, S. Pohl, Evaluating clinical accuracy of systems for self-monitoring of blood glucose, Diabetes Care 10 (5) (1987) 622–628. doi:10.2337/diacare.10.5.622.

[20] D. C. Klonoff, C. Lias, R. Vigersky, W. Clarke, J. L. Parkes, D. B. Sacks, M. S. Kirkman, B. Kovatchev, the Error Grid Panel, The surveillance error grid, J. of Diabetes Science and Technology 8 (4) (2014) 658–672. doi:10.1177/1932296814539589.

[21] A. N. Tikhonov, V. B. Glasko, Use of the regularization method in nonlinear problems, USSR Computational Math. and Math. Phys. 5 (3) (1965) 93–107. doi:10.1016/0041-5553(65)90150-3.