Aggregation of regularized rankers by means of a linear functional strategy

Sergei V. Pereverzyev

Johann Radon Institute for Computational and Applied Mathematics (RICAM)
Austrian Academy of Sciences (ÖAW)
Linz, Austria

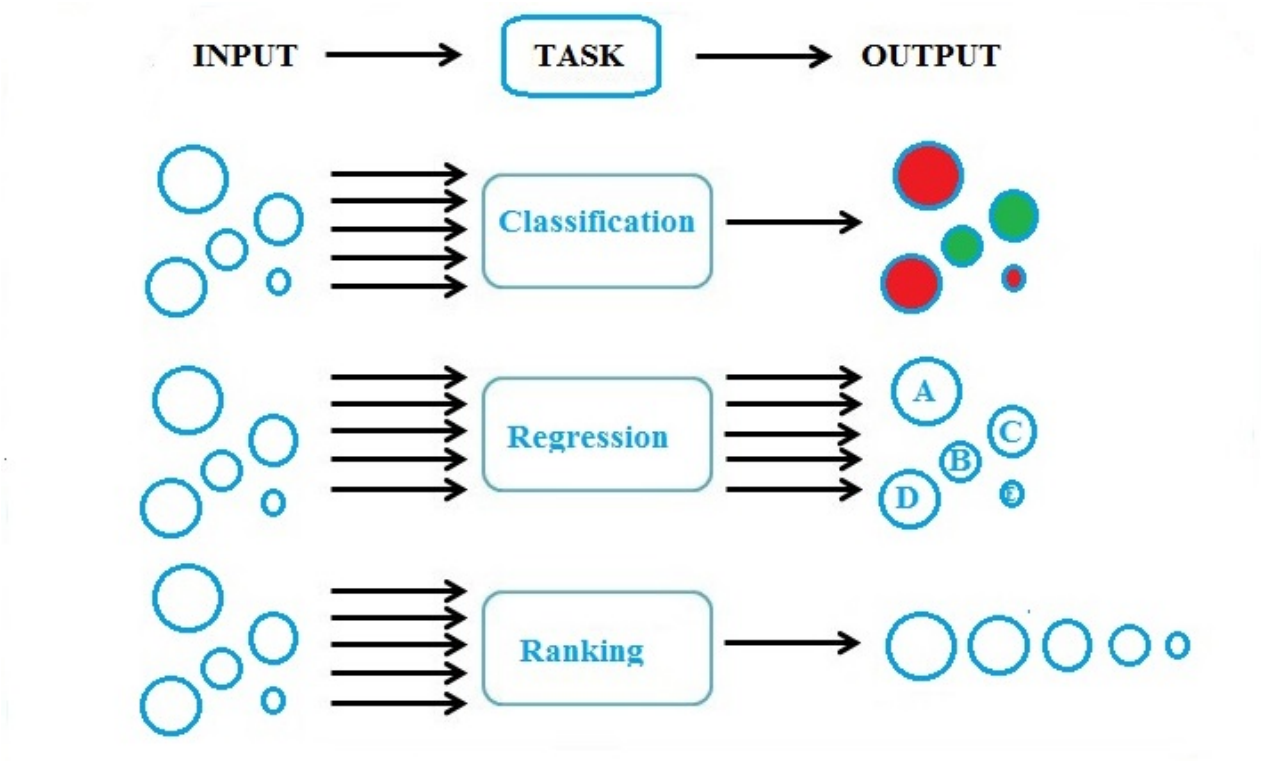
Mathematical Statistics and Inverse Problems CIRM, Marseille, February 8–12, 2016

FWF Der Wissenschaftsfonds.
HORIZON 2020 LE PROGRAMME DE RECHERCHE ET D'INNOVATION DE L'UNION EUROPÉENNE

Learning from Examples of the input-output pairs

- An input $x_i \in X \subset \mathbb{R}^d$ may cause an output $y_i \in Y \subset [-M; M]$.
- We are given a training set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z = X \times Y$ of examples of the input-output pairs.
- The problem is to learn from $\mathbf{z}$ how to predict the input-output behavior of a given black box.

Types of learning tasks



Learning ranking from examples

Herbrich et al. (2000), Crammer, Singer (2001), Freund et al. (2003), Mukherjee, Zhou (2006), Cossock, Zhang (2006), Agarwal, Niyogi (2009), Chen (2012):

1. The inputs $x \in X \subset \mathbb{R}^d$ are related to their ranks $y \in Y = [-M, M]$ through an unknown probability distribution $\rho(x, y) = \rho(y|x)\rho_X(x)$ on $Z = X \times Y$.
2. The distribution ρ is given only through a set of samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \subset Z$.

The Ranking Problem

The problem is to learn from $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ a ranking function $f = f_{\mathbf{z}} : X \to Y$ that predicts the risk of x as $f(x)$, for any $x \in X$.
A performance measure. For given true ranks y and y' of the inputs $x, x' \in X$ the value

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

is interpreted as the *magnitude-preserving least squares loss* of a ranking function f.
Then the quality of a ranking function is measured via the expected risk

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

The target functions

In the space $L_2(X, \rho_X)$ the risk $\mathcal{E}(f)$ is minimized by functions from the set

$$\mathcal{F}_\rho = \{f : f(x) = f_\rho(x) + c, c \in \mathbb{R}\},$$

where

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

is known in learning theory as *the regression function*. Then the deviation of a ranking function $f \in L_2(X, \rho_X)$ from the risk minimizers can be measured by

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \int_X \int_X \left((f_\rho(x) - f_\rho(x')) - (f(x) - f(x')) \right)^2 d\rho_X(x) d\rho_X(x')$$

An approximation in a stronger norm

Let $\mathcal{H}$ be an embedded subspace of $L_2(X, \rho_X)$, and $\mathcal{F}_\rho \subset \mathcal{H}$. Then, as in Chen (2012), the quality of a ranking function $f \in \mathcal{H}$ can be measured by the distance

$$d_{\mathcal{H}}(f, \mathcal{F}_\rho) = \inf_{g \in \mathcal{F}_\rho} \|f - g\|_{\mathcal{H}}.$$

The choice of $\mathcal{H}$ as the reproducing kernel Hilbert space (RKHS) $\mathcal{H} = \mathcal{H}_K$ associated with a kernel $K : X \times X \to \mathbb{R}$ is widely used in learning theory.
Then a ranking function may appear in minimizing a regularized risk functional

$$\mathcal{E}(f) + \alpha \|f\|^2_{\mathcal{H}_K} \to \min$$

Ranking by Lavrentiev regularization

Chen (2012) has observed that the minimizer $f = f^\alpha$ of the regularized risk functional satisfies the following equation

$$\left(\tfrac{\alpha}{2}\mathbb{I} + L_K\right) f = L_K f_\rho,$$

where the integral operator

$$L_K f(\cdot) = \int_X \int_X f(x)\left(K(x,\cdot) - K(x',\cdot)\right) d\rho_X(x) d\rho_X(x')$$

can be seen as a self-adjoint positive linear operator in $\mathcal{H}_K$. Then f^α is Lavrentiev regularized approximation to a solution of

$$L_K f = L_K f_\rho.$$

A quick overview of known efficiency estimates

- Under the assumption that the ideal input-output predictor has a Hoelder-type regularity of order r measured through the corresponding source condition in $\mathcal{H}_K$ the following is known.

Learning algorithm and regularity range	Regression Learning		Ranking Learning	
	Excess risk	Learning rate in $\mathcal{H}_K$	Excess risk	Learning rate in $\mathcal{H}_K$
Tikhonov/ Lavrentiev $r \in (0;1]$	$O(m^{-\frac{2r+1}{2r+2}})$	$O(m^{-\frac{r}{2r+2}})$	$O(m^{-\frac{r}{2r+3}})$	$O(m^{-\frac{r}{2r+3}})$
	Smale, Zhou (2007)		Chen (2012)	
General Regularization scheme $r \in (0;\infty)$	$O(m^{-\frac{2r+1}{2r+2}})$	$O(m^{-\frac{r}{2r+2}})$	?!	?!
	Bauer, Pereverzyev, Rosasco (2007)			
Spectral Regularization $r \in (0;\infty)$	$O(m^{-\frac{2r+1}{2r+2}})$	$O(m^{-\frac{r}{2r+2}})$	?!	$O(m^{-\frac{r}{2r+3}})$
	Bauer, Pereverzyev, Rosasco (2007)			Xu, Fang, Wang (2014)
Online gradient descent learning	$O(m^{-\frac{2r+1}{2r+2}})$	$O(m^{-\frac{r}{2r+2}})$	$O(m^{-\frac{2r+1}{2r+3}})$	?
	Ying, Pontil (2008)		Ying, Zhou (2015)	

Sample discretization

The regularized empirical risk minimization

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

leads to the ranking function

$$f_{\mathbf{z}}^{\alpha} = \left(\tfrac{1}{m^2}S_{\mathbf{x}}^*\mathbb{D}S_{\mathbf{x}} + \tfrac{\alpha}{2}\mathbb{I}\right)^{-1}\tfrac{1}{m^2}S_{\mathbf{x}}^*\mathbb{D}\mathbf{y},$$

that can be seen as Lavrentiev regularized approximation to a solution of discrete equation

$$\tfrac{1}{m^2}S_{\mathbf{x}}^*\mathbb{D}S_{\mathbf{x}}f = \tfrac{1}{m^2}S_{\mathbf{x}}^*\mathbb{D}\mathbf{y},$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_m)^T \in \mathbb{R}^m$, $\mathbb{D} = m\mathbb{I} - \mathbf{1}\times\mathbf{1}^T$,
$\mathbb{I}$, $\mathbf{1}$ 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 $S_{\mathbf{x}}$.

Improvements and Generalizations

Definition

$\varphi : (0, d) \to \mathbb{R}$ is called *operator monotone* (o.m.f.) on $(0, d)$ if $\forall A_i = A_i^* : X \to X$, $sp(A_i) \subset (0, d)$, $i = 1, 2$,

$$A_1 \geq A_2 \Rightarrow \varphi(A_1) \geq \varphi(A_2).$$

$$\Psi_d = \{\varphi : (0, d) \to \mathbb{R},\ \varphi \text{ is o.m.f.}, \varphi(0) = 0\}.$$

$$\Phi_d = \{\varphi : \varphi = \varphi_1 \cdot \varphi_2,\ \varphi_1 \in \Psi_d,\ \varphi_2 \text{ is Lipschitz monotone}, \varphi_2(0) = 0\}.$$

Examples

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

Improvements and Generalizations (Continuation)

$$L_K f = L_K f_\rho \qquad \longrightarrow \qquad \tfrac{1}{m^2} S_{\mathbf{x}}^* \mathbb{D} S_{\mathbf{x}} f = \tfrac{1}{m^2} S_{\mathbf{x}}^* \mathbb{D}\mathbf{y},$$

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

Definition

A regularization scheme generated by a family of functions $\{g_\alpha\}$ has a qualification p if $\exists \gamma_p > 0 : \forall \alpha > 0$

$$\sup_t t^p |1 - t g_\alpha(t)| \leq \gamma_p \alpha^p.$$

Examples

1. Lavrentiev regularization is generated by $g_\alpha(t) = \left(\frac{\alpha}{2} + t\right)^{-1}$ and has qualification $p = 1$.
2. p-times iterated Lavrentiev regularization is generated by $g_\alpha(t) = t^{-1}\left(1 - (\alpha/(\alpha + 2t))^p\right)$ and has qualification p.

Improvements and Generalizations (Continuation)

Definition (Mathé & Pereverzev, 2003)

$\varphi \in \Psi_d \cup \Phi_d$ is covered by qualification p if $\frac{t^p}{\varphi(t)}$ is non-decreasing function of $f \in [0, d]$.

Theorem (1)

Assume that $f_\rho \in Range(\varphi(L_K))$, $\varphi \in \Psi_d \cup \Phi_d$, $d > \sup_x |K(x,x)| \left(2 + 26 m^{-\frac{1}{2}} \log \frac{1}{\delta}\right)$, *and* φ *is covered by qualification of* $\{g_\alpha\}$. *Then for* $\alpha = \Theta^{-1}(m^{-1/2})$, $\Theta(t) = \varphi(t)t$, *with confidence* $1 - \delta$ *we have*

$$\inf_{g \in \mathcal{F}_\rho} \|f_{\mathbf{z}}^\alpha - g\|_{\mathcal{H}_K} = O\left(\varphi\left(\Theta^{-1}(m^{-\frac{1}{2}})\right) \log \frac{1}{\delta}\right)$$

Note. For $\varphi(t) = t^r$, $0 < r \leq 1$, we have the convergence rate in $\mathcal{H}_K$ of order $O\left(m^{-\frac{r}{2r+2}}\right)$ that improves the order $O\left(m^{-\frac{r}{2r+3}}\right)$ known previously (Chen, 2012)

Ranking by the linear functional strategy

- The usual parameter choice rules require the computation of a sequence $\{f_{\mathbf{z}}^{\alpha_i}\}$, although they select just one candidate out of such a sequence.
- The idea is to use $\{f_{\mathbf{z}}^{\alpha_i}\}_{i=1}^{I}$ as a basis for constructing a new ranking function

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

- In contrast to known aggregation procedures by Nemirovski (2000), Bunea, Tsybakov & Wegkamp (2007) or Hebiri, Loubes & Rochet (2014) no penalization or sample splitting will be used. Moreover, an aggregation will be performed not in the empirical norm, but in terms of excess risk and learning rate.

Ranking by the linear functional strategy

- The vector $\overline{c} = (\overline{c}_p)$ of the ideal coefficients for approximation in a Hilbert space $\mathcal{H} \hookrightarrow L_2(X, \rho_X)$ minimizes the norm

$$\|f_\rho - \sum_p c_p f_{\mathbf{z}}^{\alpha_p}\|_{\mathcal{H}} \to \min$$

  and 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 $\overline{g} = \left(\langle f_\rho, f_{\mathbf{z}}^{\alpha_p}\rangle_{\mathcal{H}}\right)_{p\in\Pi}$

- The regularization theory tells (see, e.g. Goldenshluger, Pereverzev (2000), Bauer, Mathé, Pereverzev (2007), Lu, Pereverzev (2013)) that the values of linear bounded functionals $\langle f_{\mathbf{z}}^{\alpha_p}, \cdot\rangle_{\mathcal{H}}$ at f_ρ can be estimated by the so-called *linear functional strategy* much more accurately than f_ρ in $\mathcal{H}$.

Convergence rate of the discretized LFS

Known results on the linear functional strategy are obtained under the assumption that equation operators, such as L_K, are accessible, which is not the case in Ranking.

Theorem (2)

Assume that $f_\rho \in Range(\varphi(L_K))$, $l \in Range(\psi(L_K))$, $\varphi \in \Phi_d \cup \Psi_d$, $\psi \in \Psi_d$, *and* d *meets the condition of Theorem 1. If the product* $\varphi \cdot \psi$ *is covered by the qualification of* $\{g_\alpha\}$ *then for* $\alpha = \Theta^{-1}(m^{-1/2})$, $\Theta(t) = \varphi(t)t$, *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.

Note due to Mathé, Hofmann (2008) for any $l \in \mathcal{H}_K$ there is always a function ψ such that $l \in Range(\psi(L_K))$. Then in view of Examples of o.m.f. the assumption that ψ is o.m.f. is not a real restriction.

A by-product result on the excess risk

Theorem (3)

Assume the conditions of Theorem 1. Then for $\alpha = \Theta^{-1}(m^{-1/2})$ *with confidence* $1-\delta$ *we have*

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

Sketch of the proof: At first we observe that

$$\mathcal{E}(f_{\mathbf{z}}^{\alpha}) - \mathcal{E}(f_{\rho}) = \langle L_K(f_{\mathbf{z}}^{\alpha} - f_{\rho}), (f_{\mathbf{z}}^{\alpha} - f_{\rho})\rangle_{\mathcal{H}_K} = \|L_K^{\frac{1}{2}}(f_{\mathbf{z}}^{\alpha} - f_{\rho})\|_{\mathcal{H}_K}^2$$

Then using Theorem 2 with $\psi(t) = \sqrt{t}$ we obtain $\|L_K^{\frac{1}{2}}(f_{\mathbf{z}}^{\alpha} - f_{\rho})\|_{\mathcal{H}_K} =$

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

Note For $\varphi(t) = t^r$ Theorem 3 gives an estimation of the excess risk of order $O\left(m^{-\frac{2r+1}{2r+2}}\right)$ that improves the order $O\left(m^{-\frac{r}{2r+3}}\right)$ given by Chen (2012).

Ranking by the LFS (Continuation)

Due to Theorem 2 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 $\tilde{g} = \left(\langle f_{\mathbf{z}}^{\alpha_{l_p}}, f_{\mathbf{z}}^{\alpha_p}\rangle_{\mathcal{H}_K}\right)_{p\in\Pi}$ 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. In our numerical tests we take α_p, $p = 0, 152, 200$, from the sequence $\{\alpha_i = (0.97)^i, i = 0, 1, \ldots, 200\}$, because such α_p have three different orders of magnitude $10^0, 10^{-2}, 10^{-3}$.

Ranking by the LFS (Continuation)

- Theorem 2 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. 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 $\{\alpha_i = (0.97)^i, i = 0, 1, \ldots, 200\}$

Ranking by the linear functional strategy in $\mathcal{H}_K$

Thus, 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 1 with confidence $1-\delta$ it holds

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

This means that we can effectively construct a ranking function from $span\{f_{\mathbf{z}}^{\alpha_p}, p \in \Pi\}$, 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.

Ranking by the linear functional strategy in $L_2(X, \rho_X)$

In the case $\mathcal{H} = L_2(X, \rho_X)$ 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 (1)

Under the conditions of Theorem 1 with condifence $1-\delta$ *it holds*

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

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

Note In the space $\mathcal{H} = L_2(X, \rho_X)$ the estimation of the values of linear bounded functional $\langle f_{\mathbf{z}}^{\alpha_p}, \cdot\rangle_{\mathcal{H}}$ at $\overline{f_\rho}$ is not an ill-posed problem, such that no regularization is required.

Ranking by the LFS in $L_2(X, \rho_X)$ (continuation)

Proposition (2)

Let $\tilde{G} = \left(m^{-1}\langle S_{\mathbf{x}} f_{\mathbf{z}}^{\alpha_p}, S_{\mathbf{x}} f_{\mathbf{z}}^{\alpha_j}\rangle_{\mathbb{R}^m}\right)_{p,j\in\Pi}$, $\tilde{g} = \left(m^{-2}\langle \mathbb{D}\mathbf{y}, S_{\mathbf{x}} f_{\mathbf{z}}^{\alpha_p}\rangle_{\mathbb{R}^m}\right)_{p\in\Pi}$. *Then under the condition of Theorem 1 with confidence* $1-\delta$ *in holds*

$$\|\tilde{G} - G\|_{\mathbb{R}^q\to\mathbb{R}^q} = O(m^{-\frac{1}{2}}\log\tfrac{1}{\delta}), \quad \|\tilde{g} - \overline{g}\|_{\mathbb{R}^q} = O(m^{-\frac{1}{2}}\log\tfrac{1}{\delta}).$$

Theorem (4)

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

$$\|\overline{f_\rho} - f_{\mathbf{z}}\|_{L_2(X,\rho_X)} = \min_{c_p}\|\overline{f_\rho} - \sum_{p\in\Pi}\tilde{c}_p f_{\mathbf{z}}^{\alpha_p}\|_{L_2(X,\rho_X)} + O(m^{-\frac{1}{2}}\log\tfrac{1}{\delta}).$$

This means that in $L_2(X, \rho_X)$ the distance of $f_{\mathbf{z}}$ to a risk minimizer differs from the best approximation by a quantity of parametric rate order $O(m^{-\frac{1}{2}})$.

Numerical example, Micchelli & Pontil (2005)

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

$$y = f_\rho(x) + \epsilon, \ \epsilon \in [-0.02, 0.02]$$

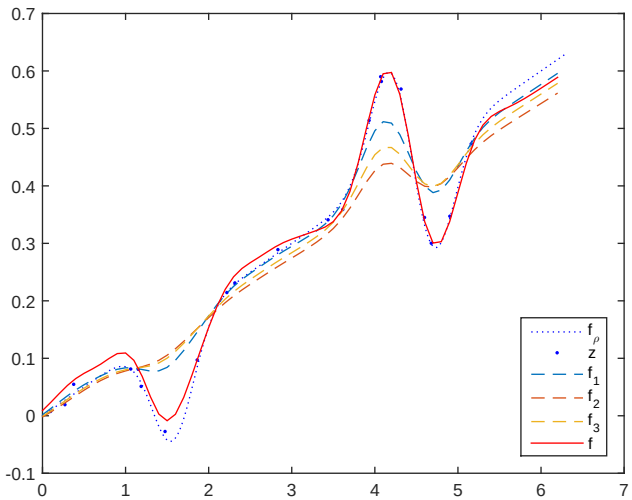


Figure: The target function f_ρ (red line), training set $\mathbf{z}$, $f_{\mathbf{z}}^{\alpha}$ for $\alpha = 0.1, 0.2, 0.3$, and $f_{\mathbf{z}}$

$$K(x, x') = e^{-8(x-x')^2} + x \cdot x'$$

Function	Fraction of misranked pairs	Mean squared error
$f_1 = f_{\mathbf{z}}^{0.1}$	6.04%	0.0020
$f_2 = f_{\mathbf{z}}^{0.2}$	7.58%	0.0032
$f_3 = f_{\mathbf{z}}^{0.3}$	7.89%	0.0041
$f_{\mathbf{z}}$	1.18%	0.00032

LFS for the prediction of Nocturnal Hypoglycemia (NH)

- NH (a low BG-concentration during the sleep period) is the most common and particular worrisome hypoglycemia in individuals with diabetes.
- One of the first method for predicting NH was proposed by Whincup and Milner (1987). 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 by Whincup and Milner (1987) 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-predict.

Linear functional strategy in NH-prediction

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

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

where $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ is a training set 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.

- Following the ranking by the linear functional strategy in $L_2(X, \rho_X)$. 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

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

and

$$S_x f_{a_p} = (f_{a_p}(x_1), f_{a_p}(x_2), \ldots, f_{a_p}(x_m)), \ y = (y_1, y_2, \ldots, y_m).$$

Test on clinical data

- 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 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 $\mathbf{z}$.
- Following Whincup and Milner (1987) 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.

Comparative Performance of NH-predictors

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**

Test on clinical data

- As it can be seen from the table the ranking function f_{a_3}, that was suggested by Whincup and Milner (1987) 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 by Whincup and Milner (1987), 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.

Nocturnal Hypoglycemia risk

An extension of the number of possible outputs (risks) of the ranking function $f_\rho(x)$ can provide diabetes patients with more information on their health state. For example, instead of using classification $\{-1, 1\}$ (similar to saying NO or YES), we may set the following scenarios and the corresponding ranks:

$$f_\rho(x) = \begin{cases} 2, & \text{very high risk of NH,} \\ 1, & \text{high risk of NH,} \\ 0, & \text{moderate risk of NH,} \\ -1, & \text{low risk of NH,} \\ -2, & \text{no risk of NH.} \end{cases}$$

Predictors based on low blood glucose index (LBGI)

- LGBI was introduced originally in (Kovatchev et. al. 1998) for measuring the risk of severe hypoglycemia (SH), based on SBGM.
- LBGI takes nonnegative continuous values.
- It was observed that LBGI was significantly higher on the day prior to an SH.

Method 1: Calculate the LBGI based on 4 measurements during a day.
Method 2: Calculate the LGBI based on the latest before bed BG-measurement (analog of the method by Whincup and Milner (1987)).

Assign the ranks similar to ACCU-CHEK® Connect

$$f_{lbgi}(x) = \begin{cases} 2, & lbgi \geq 5, \\ 1, & 2.5 \leq lbgi < 5, \\ 0, & 1 \leq lbgi < 2.5, \\ -1, & 0.5 \leq lbgi < 1, \\ -2, & lbgi < 0.5. \end{cases}$$

Aggregation of predictors based on LBGI

We aggregate these two predictors and approximate $\bar{f}_\rho(x)$ by

$$f_{\mathbf{z}} = \tilde{c}_1 f_{lbgi_1}(x) + \tilde{c}_2 f_{lbgi_4}(x).$$

To measure the performance of 3 predictors in terms of SE, SP, PPV, NPV and F1-score we set the following classification condition:

Each non-negative output of predictors (0, 1 or 2 for LBGI predictors) we interpret as YES (set predicted value as 1), each negative as NO (-1).

Our test were performed on the same set of data and in the same way as for ranking functions from Whincup and Milner (1987).

Comparative Performance of NH-predictors

Ranking function	SE (%)	SP(%)	PPV (%)	NPV (%)	F_1	Pairwise Misranking
f_ρ	1	1	1	1	1	
f_{lbgi_4}	65.69	95.63	84.74	88.31	0.7378	0.4452
f_{lbgi_1}	55.56	97.29	88.29	85.63	0.6789	0.5755
f_z	79.91	91.28	78.63	92.60	**0.7870**	**0.3884**

Application to the data from AMMODIT

- Using a dataset of 150 nights, collected within the EU-FP7 Project DIAdvisor we aggregate in f_z the NH-predictors known in the literature.
- Within the EU-Horizon 2020-project AMMODIT the trained predictor f_z with **fixed coefficients** $\tilde{c}$ has been implemented as a **Smartphone App** and tested on another database consisting of 476 nights (different patients, children), collected in a Ukrainian hospital.
- Additionally, the predictor f_z has been tested on data (182 day records) of a **single patient**.

	SE (%)	SP(%)	PPV (%)	NPV (%)	F_1-score
Ukrainian dataset	77.03	81.89	78.80	80.31	0.7790
Single patient	69.23	79.23	57.14	86.55	0.6261
State of the art (HypoMon)	72.73	68.29	38.10	90.32	0.5000

DIApvisor vs HypoMon