Opuscula Math. 36, no. 5 (2016), 671–679
http://dx.doi.org/10.7494/OpMath.2016.36.5.671 Opuscula Mathematica

ON A DENSE MINIMIZER OF EMPIRICAL RISK IN INVERSE PROBLEMS

Jacek Podlewski and Zbigniew Szkutnik

Communicated by Andrzej Kozek

Abstract. Properties of estimators of a functional parameter in an inverse problem setup are studied. We focus on estimators obtained through dense minimization (as opposed to minimization over δ-nets) of suitably defined empirical risk. At the cost of imposition of a sort of local finite-dimensionality assumption, we fill some gaps in the proofs of results published by Klemelä and Mammen [Ann. Statist. 38 (2010), 482–511]. We also give examples of functional classes that satisfy the modified assumptions.

Keywords: inverse problem, empirical risk minimization.

Mathematics Subject Classification: 62G07.

1. INTRODUCTION

Following Klemelä and Mammen (cf., [8, 9]), we consider estimation of a function $f : \mathbb{R}^d \to \mathbb{R}$ when an i.i.d. sample $Y_1, \ldots, Y_n$ from a density Af is observed, where Y_i's take values in a linear measure space $(\mathcal{Y}, \Sigma, \mu)$, the density Af is taken w.r.t. μ, $A : L_2(\mathbb{R}^d, \lambda^d) \to L_2(\mathcal{Y}, \mu)$ is a linear, invertible operator, λ^d stands for the standard d-dimesional Lebesgue measure and f is assumed to belong to some class $\mathcal{F} \subset L_2(\mathbb{R}^d, \lambda^d)$. This is a form of a statistical inverse problem. Following [4] and [9], we define the empirical risk functional as

$$\gamma_n(g) = -\frac{2}{n}\sum_{i=1}^{n}(Qg)(Y_i) + \|g\|_2^2,$$

where Q is the adjoint of the inverse of A and $\|\cdot\|_2$ stands for the L_2-norm, and study estimators defined as minimizers of this functional over $\mathcal{F}$. They are called dense minimizers, as opposed to the minimizers over a δ-net in $\mathcal{F}$. It is easily seen that with $\gamma(Y, g) = -2(Qg)(Y) + \|g\|_2^2$ one has $\mathbb{E}\gamma(Y_1, g) = \|g - f\|_2^2 - \|f\|_2^2$ and, consequently, $f = \operatorname{argmin}_g \mathbb{E}\gamma(Y_1, g)$. This means that $\gamma(Y, g)$ is a contrast function, $\gamma_n(\cdot)$ may be

viewed as an empirical contrast functional and its minimizer over $\mathcal{F}$ as a minimum contrast estimator with the corresponding natural L_2-estimation loss (cf., e.g., [1]). From the statistical learning theory point of view, $\mathbb{E}\gamma(Y_1, g)$ is the risk functional and $\gamma_n(\cdot)$ is the empirical L_2-risk functional (cf., e.g., [14, p. 18]).

Both dense and δ-net minimizers have been studied in [9] and claimed to be rate-minimax estimators under some regularity assumptions. However, the proofs of the results for the dense minimizer seem to have some gaps. We fix the problems under somewhat strengthened assumptions.

Let us recall that obtaining optimal estimators as elements of suitably constructed δ-nets has a long tradition dating back to [11]. The regularizing effect of δ-net discretizations is known to be important in non-parametric, especially inverse, problems, if the estimators are obtained as minimizers of some functionals over infinite-dimensional parameter spaces. Minimax convergence rates have been proved, e.g., for non-parametric minimum distance δ-net estimators in the L_1-setup in [15] and [12], and in a general, loss-based, or minimum contrast formulation in [13]. On the other hand, it has been shown in [1] that, in a quite general setup, (dense) minimum contrast estimators may achieve minimax convergence rate as long as the entropy function is regular enough and not too large. If the entropy function becomes too large, suboptimal rates can be expected. Nevertheless, the fact that dense minimizers in [9] should achieve minimax convergence rates in the inverse problem setup without the regularizing effect of δ-nets discretization, and with only weak entropy restrictions, seems somewhat surprising. Trying to fix the gap in the original proof, we had to impose a stronger condition on the structure of the parameter set – in fact, a sort of local finite-dimensionality condition.

Main theoretical results, i.e., the corrected versions of results from [8,9] are given in Section 2, along with a detailed discussion of changes made to their original statements. In Section 3, some examples of functional classes that fulfill the modified assumptions are given and the relation between δ-net minimizers and dense minimizers is shortly discussed.

2. RESULTS

In this section, we discuss the necessity of fixing some gaps in the proofs of Theorem 4 in [9] and of Lemma 5 in [8] and provide their corrected versions: Theorem 2.1 and Lemma 2.2, respectively.

Let $\mathcal{F} \subset L_2(\mathbb{R}^d)$. For $\delta > 0$, a set $\mathcal{F}_\delta = \{(g_j^L, g_j^U) : j = 1, 2, \dots N_\delta\}$ of pairs of L_2-functions is called a δ-bracketing of $\mathcal{F}$ if:

1. $\|g_j^U - g_j^L\|_2 \leq \delta$ for $j = 1, \dots, N_\delta$,
2. for any $g \in \mathcal{F}$, there exists $j = j(g)$ such that $g_j^L \leq g \leq g_j^U$.

We will assume that N_δ is finite for all $\delta > 0$. Following [9], denote

$$\mathcal{F}_\delta^L = \{g_j^L : j = 1, \dots, N_\delta\}, \quad \mathcal{F}_\delta^U = \{g_j^U : j = 1, \dots, N_\delta\}$$

and, for $Q = (A^{-1})^*$, define

$$\varrho\left(Q, \mathcal{F}_\delta^L, \mathcal{F}_\delta^U\right) = \max_{g^L \in \mathcal{F}_\delta^L, g^U \in \mathcal{F}_\delta^U} \left\{ \frac{\|Q(g^U - g^L)\|_2}{\|g^U - g^L\|_2} \right\},$$

$$\varrho\left(Q, \mathcal{F}_\delta^L, \mathcal{F}_{2\delta}^L\right) = \max_{f \in \mathcal{F}_\delta^L, g \in \mathcal{F}_{2\delta}^L} \left\{ \frac{\|Q(f - g)\|_2}{\|f - g\|_2} \right\},$$

and

$$\varrho_{den}\left(Q, \mathcal{F}_\delta\right) = \max\left\{ \varrho\left(Q, \mathcal{F}_\delta^L, \mathcal{F}_\delta^U\right), \varrho\left(Q, \mathcal{F}_\delta^L, \mathcal{F}_{2\delta}^L\right) \right\}.$$

Let $B_2 = \sup_{f \in \mathcal{F}} \|f\|_2$. For $r \in (0, B_2]$, define the entropy integral

$$G(r) = \int_0^r \rho_{den}(Q, \mathcal{F}_u) \sqrt{\log(\#\mathcal{F}_u)} du,$$

where $\#\mathcal{F}_u$ is the cardinality of $\mathcal{F}_u$ and $\log(\#\mathcal{F}_u)$ is called u-entropy with bracketing of $\mathcal{F}$.

Define the estimator $\hat{f}$ as the minimizer of the empirical risk functional over $\mathcal{F}$, up to some precision ϵ_n, i.e.,

$$\gamma_n(\hat{f}) \leq \inf_{g \in \mathcal{F}} \gamma_n(g) + \epsilon_n$$

for some $\epsilon_n > 0$. The following theorem is a corrected version of Theorem 4 in [9]. There is one important change with respect to the original version of the theorem: the additional assumption 8. Some modifications will also be needed in the formulation and in the proof of a main technical lemma used in the proof of the theorem.

Theorem 2.1. *Assume that:*

1. $\sup_{f \in \mathcal{F}} \|Af\|_\infty \leq B_\infty$ *for some constant* B_∞.
2. $\sup_{g \in \mathcal{F}_\delta^L \cup \mathcal{F}_\delta^U} \|Qg\|_\infty \leq B_\infty$ *for some constant* B_∞.
3. *The mapping* $\delta \to \varrho_{den}\left(Q, \mathcal{F}_\delta\right) \sqrt{\log\left(\#\mathcal{F}_\delta\right)}$ *is decreasing on* $(0, B_2]$.
4. $G(B_2) < \infty$.
5. $\varrho_{den}\left(Q, \mathcal{F}_\delta\right) = c\delta^{-a}$ *for some* $c > 0$ *and* $a \in [0, 1)$.
6. Q *preserves positivity, i.e.* $g \geq 0$ *implies that* $Qg \geq 0$.
7. $G(\delta)/\delta^2$ *is decreasing on* $(0, B_2]$ *and* $\lim_{\delta \to 0^+} G(\delta)\delta^{a-1} = \infty$.
8. *There exist constants* D *and* r *such that for each ball* $K_\delta \subset L_2(\mathbb{R}^d)$ *with radius* $\delta < r$ *there exists a* δ*-bracketing* $(K_\delta \cap \mathcal{F})_\delta$ *of* $K_\delta \cap \mathcal{F}$ *such that* $\#(K_\delta \cap \mathcal{F})_\delta \leq D$.

If ψ_n *is a sequence such that* $\psi_n^2 \geq Cn^{-1/2}G\left(\psi_n\right)$ *and* $\lim_{n \to \infty} n\psi_n^{2(1+a)} = \infty$ *for some* $C > 0$*, then, for* n *large enough and some constant* $C' > 0$*,*

$$\sup_{f \in \mathcal{F}} \mathbb{E}\|\hat{f} - f\|_2^2 \leq C'\left(\psi_n^2 + \epsilon_n\right).$$

Remarks on the proof of Theorem 2.1. Technically, the reason for imposing assumption 8 is as follows. The proof of Theorem 4 in [9] goes along the same lines as the proof of Theorem 3 in [9] up to step (79) in page 507. This means that the peeling device is used to upper bound P_{sup}, defined in formula (75), with the series given in (77). Then, Lemma 4 is used for each term of the series, which eventually majorizes the series of interest by a convergent geometrical series and, finally, by an exponential function that is further integrated to produce a desired constant. Clearly, any application of Lemma 4 has to use the same radius R in both exponential terms and in the coefficient preceeding the second term in the thesis of Lemma 4. In the proof of Theorem 4, however, the authors substitute $R = \sqrt{b_j}$ (slightly more than the radius of the jth peeling layer) in the exponential terms, but $R = B_2$ in the coefficient. This may be correct for large j, but for small j the coefficient should have the form $2 \cdot \#\{g \in \mathcal{F} : \|g - f\|_2 \leq \sqrt{b_j}\}_{\sqrt{b_j}}$, i.e., twice the cardinality of $\sqrt{b_j}$-bracketing of the (subset of) $\sqrt{b_j}$-ball. This quantity has to be bounded by an absolute constant for the proof of Theorem 4, because, e.g., b_0 converges to zero as n tends to infinity, which may potentially give exploding number of $\sqrt{b_0}$-brackets needed to cover $\{g \in \mathcal{F} : \|g - f\|_2 \leq \sqrt{b_0}\}$. Our additional assumption 8 fixes that. □

Some changes were also needed to fix a gap in the proof of Lemma 5 in [8]. In [9] the same lemma has number 4 and it is a crucial tool in the proof of Theorem 4 (and of our Theorem 2.1), providing exponential bounds for the tail probabilities of a centered empirical process. For better readability, we recall (a slightly modified version of) the lemma.

Let

$$\nu_n(g) = n^{-1} \sum_{i=1}^{n} g(Y_i) - \int_{\mathcal{Y}} g(y)(Af)(y) d\mu(y)$$

be the centered empirical process.

Lemma 2.2. *Let $\mathcal{G} \subset L_2(\mathbb{R}^d, \lambda^d)$ and $R = \sup_{g \in \mathcal{G}} \|g\|_2$. Assume that $\|Af\|_\infty \leq B_\infty$ for the true function f and that conditions 2-6 of Theorem 2.1 hold true with $\mathcal{F}$ replaced with $\mathcal{G}$ and B_2 replaced with R. Then, for all ξ such that $\xi \geq n^{-1/2}\tilde{G}(R)$, where*

$$\tilde{G}(R) = (1 - 2^{a-1})^{-1} \sqrt{B_\infty \left(9^2 + 96 \cdot 2^{-2a}\right)} \times \max\left\{24\sqrt{2}G(R), \frac{4}{\log 2}\left[C_a + (1-a)^{-3/2}\Gamma(3/2)\right] cR^{1-a}\right\}$$

and $C_a = \sqrt{2}\left(1 - 2^{a-1}\right)^{-1/2}$, one has

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} \nu_n(Qg) \geq \xi\right) \leq 4 \exp\left(-\frac{n\xi^2 C'}{B_\infty c^2 R^{2-2a}}\right) + 2\#\mathcal{G}_R \exp\left\{-\frac{n^2\xi^2}{72}\left(B_\infty c^2 R^{2-2a} + \frac{2}{9}\xi B'_\infty\right)^{-1}\right\},$$

for some constant $C' > 0$.

Remarks on the proof of Lemma 2.2. Notice that assumption 6 of positivity-preservation property of Q appears in the original statement of Theorem 2.1 without being used explicitly in its proof, but it is missing in the original formulation of Lemma 2.2, although it is clearly exploited in the proof of the lemma (see (2.1) in the discussion below).

Our version of the lemma also slightly differs from the original one in the form of some constants. The constant C_a is missing in the original definition of $\tilde{G}(R)$, and the denominator in the exponent in the last term of the thesis equals 12 rather than 72. We were only able to prove Lemma 2.2 in its current form. It should be stressed, however, that the changes affect neither the application of the lemma, nor the conclusions of the theorem.

We have also added the factor $(1-2^{a-1})^{-1}$ in the definition of $\tilde{G}(R)$, because of the following reason. In the proof of the lemma, with $R = \sup_{g\in\mathcal{G}} \|g\|_2$ and $R_k = 2^{-k}R$, one considers δ-bracketings $\mathcal{G}_\delta$ of a class $\mathcal{G} \subset L_2(\mathbb{R}^d)$ with $\delta = R_k$. Let $(h_g^{k,L}, h_g^{k,U})$ be a member of the bracketing net $\mathcal{G}_{R_k}$ such that $h_g^{k,L} \le g \le h_g^{k,U}$ and define $\Delta_g^k = h_g^{k,U} - h_g^{k,L}$. For the displayed formula directly preceeding (128) in page 47 of [8] to hold true, one needs

$$\Delta_g^k \le \Delta_g^{k-1}, \tag{2.1}$$

as that implies $Q\Delta_g^k \le Q\Delta_g^{k-1}$. However, inequality (2.1) need not be true, unless the R_k-bracketings are nested, which is not guaranteed, but may be enforced in a more-or-less standard way at the price of increasing the number of brackets and modifying some constants.

For the bracketing $\{(h_i^{k,L}, h_i^{k,U}) : i = 1, \dots, N_k\}$ define disjoint classes

$$\mathcal{G}_i^k = \left\{ g \in \mathcal{G} : i = \min\{j : h_j^{k,L} \le g \le h_j^{k,U}\} \right\}$$

and note that $\mathcal{G} = \bigcup_i \mathcal{G}_i^k$. Assume that the bracketing is minimal so that all $\mathcal{G}_i^k$ are nonempty. In general, $\{\mathcal{G}_i^{k+1} : i = 1, \dots, N_{k+1}\}$ does not have to be a subpartition of the partition $\{\mathcal{G}_i^k : i = 1, \dots, N_k\}$. However, the classes $\mathcal{G}_{ij}^{k+1} = \mathcal{G}_i^k \cap \mathcal{G}_j^{k+1}$ with $i = 1, \dots, N_k$ and $j = 1, \dots, N_{k+1}$ define a partition of $\mathcal{G}$ into at most $N_k \cdot N_{k+1}$ subsets (some may be empty) corresponding to the bracketing

$$\left\{ \left(h_j^{k+1,L} \vee h_i^{k,L}, h_j^{k+1,U} \wedge h_i^{k,U} \right) : i = 1, \dots, N_k, j = 1, \dots, N_{k+1} \right\}$$

in which only those brackets are kept for which $h_j^{k+1,L} \vee h_i^{k,L} \le h_j^{k+1,U} \wedge h_i^{k,U}$. In effect, any sequence of bracketings with cardinalities $N_1, N_2, \dots$ can be transformed into a nested sequence of bracketings such that the cardinality of the kth bracketing does not exceed $N_1 \cdot \ldots \cdot N_k$ and for the corresponding R_k-bracketing entropy one obtains $H_k = \log(N_1 \cdot \ldots \cdot N_k)$.

The entropy integral is used in [8] in the proof of Lemma 5 only to obtain the last displayed inequality in the proof in page 49. After transforming an arbitrary sequence of bracketings to a nested one, one has, using $\rho_{den}(Q, \mathcal{G}_{R_k}) = cR_k^{-a}$ with $a \in [0, 1)$,

$$\sum_{k=1}^{\infty} R_k \, \rho_{den}(Q, \mathcal{G}_{R_k}) H_k^{1/2} = c \sum_{k=1}^{\infty} R_k^{1-a} \Big(\sum_{j=1}^{k} \log N_j \Big)^{1/2}$$
$$\leq cR^{1-a} \sum_{k=1}^{\infty} 2^{-k(1-a)} \sum_{j=1}^{k} (\log N_j)^{1/2} = cR^{1-a} \sum_{j=1}^{\infty} (\log N_j)^{1/2} \sum_{k=j}^{\infty} (2^{a-1})^k$$
$$= \frac{2}{1 - 2^{a-1}} \sum_{j=1}^{\infty} \frac{R}{2^{j+1}} \rho_{den}(Q, \mathcal{G}_{R_j}) \sqrt{\log(\#\mathcal{G}_{R_j})} \leq \frac{2}{1 - 2^{a-1}} G(R),$$

which is sufficient for completion of the proof of the lemma only with our modified definition of $\tilde{G}(R)$. □

3. FUNCTIONAL SPACES MEETING THE ASSUMPTIONS

The property postulated in assumption 8 in Theorem 2.1 is a bracketing analogue of, so-called, finite doubling dimension, or finite Assouad dimension (cf., e.g., [6] or [7, p. 81]). Although it is a restrictive assumption, we were not able to complete the proof without imposing it.

To give an example of a class that satisfies assumption 8, let $B_{p,q}^{\alpha} := B_{p,q}^{\alpha}(L_p[0,1])$ be the Besov space of functions on $[0,1]$, as defined, e.g., in [5, p. 54]. Notice that $B_{p,\infty}^{\alpha}(L_p[0,1]) \subset L_2([0,1])$ for $p \geq 2$.

Proposition 3.1. *For $d = 1$, $p \geq 2$ and $\alpha > 1/p$, assumption 8 in Theorem 2.1 is satisfied for any subset $\mathcal{F}$ of $B_{p,\infty}^{\alpha}$ for which*

$$M := \sup \left\{ \frac{\|f - g\|_{B_{p,\infty}^{\alpha}}}{\|f - g\|_2} : f, g \in \mathcal{F}, f \neq g \right\} < \infty. \tag{3.1}$$

Proof. Given a function class $\mathcal{F}$, let $H_{\delta,\infty}(\mathcal{F})$ denote the log-cardinality of the minimal δ-net with respect to the norm $\|\cdot\|_{\infty}$. For a Besov R-ball, defined as $K_{p,\alpha,\infty}(0,R) = \Big\{ f \in B_{p,\infty}^{\alpha} : \|f\|_{B_{p,\infty}^{\alpha}} \leq R \Big\}$, we have the following: if $\alpha > 1/p$, then for any $\delta > 0$

$$H_{\delta,\infty}(K_{p,\alpha,\infty}(0,R)) \leq C(1 + (R/\delta)^{1/\alpha}), \tag{3.2}$$

where C is a constant depending only on α and p. This follows as a special case of Proposition 2 in [2], or Theorem 1.1 in [3]. For any $\delta > 0$ and $K(\phi, \delta)$ - a δ-ball in $L_2(\mathbb{R})$ with centre ϕ, the set $\mathcal{F} \cap K(\phi, \delta)$ has L_2-diameter bounded by 2δ, so there exists $\phi_0 \in \mathcal{F}$ such that $\mathcal{F} \cap K(\phi, \delta) \subset \mathcal{F} \cap K(\phi_0, 2\delta)$. For every $f \in \mathcal{F} \cap K(\phi_0, 2\delta)$, we have $\|f - \phi_0\|_{B_{p,\infty}^{\alpha}} \leq 2M\delta$, so $f - \phi_0 \in K_{p,\alpha,\infty}(0, 2M\delta)$ and there exists a pair (h^L, h^U) from a δ-bracketing of $K_{p,\alpha,\infty}(0, 2M\delta)$ such that $h^L + \phi_0 < f < h^U + \phi_0$. Therefore,

$$\log \left[\# \left(\mathcal{F} \cap K(\phi, \delta) \right)_{\delta} \right] \leq \log \left[\# \left(K_{p,\alpha,\infty}(0, 2M\delta) \right)_{\delta} \right].$$

Because $\log(\#\mathcal{G}_\delta) \le H_{\delta/2,\infty}(\mathcal{G})$ for any function class $\mathcal{G}$ and any norm dominated by the sup-norm (see, e.g., Lemma 9.22 in [10]), one can use (3.2) to obtain

$$\log\left[\#\left(\mathcal{F} \cap K(\phi,\delta)\right)_\delta\right] \le C(1+(4M)^{1/\alpha}),$$

so the class $\mathcal{F}$ satisfies assumption 8 in Theorem 2.1. Since entropy bounds analogous to (3.2) hold for Besov spaces of funcions defined on domains other than $[0,1]$ (cf., [3]), the presented reasoning can easily be generalized into multivariate setting. □

As a more specific example, we give the following proposition.

Proposition 3.2. *Set $r = [\alpha]+1$, where $[\alpha]$ stands for the integer part of α. In the Sobolev space $W_2^r \subset B^\alpha_{2,\infty}$ pick an L_2-orthogonal system $\{\phi_j\}$ such that $\|\phi_j\|_2 \le 1$ and $\|\phi_j^{(r)}\|_2 \le L, j = 1,2,\dots$ for some constant L and define $\mathcal{F}^{(r)}_{L,N}$ to be the set of N-element linear combinations of functions from $\{\phi_j\}$. Then $\mathcal{F}^{(r)}_{L,N}$ satisfies condition* (3.1) *with $p = 2$.*

Proof. Recall that the Besov norm is defined as $\|f\|_{B^\alpha_{p,\infty}} := \|f\|_p + |f|_{B^\alpha_{p,q}}$, with the seminorm $|f|_{B^\alpha_{p,\infty}} = \sup_{t>0} t^{-\alpha}\omega_r(f,t)$ and the r-th modulus of smoothness $\omega_r(f,t)$, defined as in, e.g., [5]. For any $f, g \in \mathcal{F}^{(r)}_{L,N}$, the function $f-g$ can be expressed as $\sum_{i=j_1}^{j_{2N}} \beta_i\phi_i$ for some $j_1,\dots,j_{2N}$. Without loss of generality assume $j_i = i$ for $i = 1,\dots,2N$. We have then:

$$\begin{aligned}
|f-g|^2_{B^\alpha_{2,\infty}} &\le \|(f-g)^{(r)}\|_2^2 = \left\|\sum_{i=1}^{2N}\beta_i\phi_i^{(r)}\right\|_2^2 \\
&= \sum_{i=1}^{2N}\beta_i^2\left\|\phi_i^{(r)}\right\|_2^2 + \sum_{i\ne j}\langle\beta_i\phi_i^{(r)},\beta_j\phi_j^{(r)}\rangle_{L_2} \\
&\le \sum_{i=1}^{2N}\beta_i^2\left\|\phi_i^{(r)}\right\|_2^2 + \sum_{i\ne j}\beta_i\beta_j\left\|\phi_i^{(r)}\right\|_2\left\|\phi_j^{(r)}\right\|_2 \\
&\le \sum_{i=1}^{2N}\beta_i^2\left\|\phi_i^{(r)}\right\|_2^2 + \frac{1}{2}\sum_{i\ne j}\left(\beta_i^2\left\|\phi_i^{(r)}\right\|_2^2 + \beta_j^2\left\|\phi_j^{(r)}\right\|_2^2\right) \\
&\le (2N+1)\sum_{i=1}^{2N}\beta_i^2\left\|\phi_i^{(r)}\right\|_2^2 \le (2N+1)L\sum_{i=1}^{2N}\beta_i^2 \\
&\le (2N+1)L\|f-g\|_2^2.
\end{aligned}$$

The reasoning above first used the fact that the Besov seminorm $|\cdot|_{B^\alpha_{2,\infty}}$ is bounded above by the Sobolev seminorm (because $\omega_r(f,t) \le t^r\|f^{(r)}\|_2$, see inequality (7.12) in [5], and $r > \alpha$), followed by an application of the Cauchy-Schwarz inequality and the orthogonality of $\{\phi_i\}$. Hence, $\|f-g\|_{B^\alpha_{2,\infty}} \le (1+\sqrt{(2N+1)L})\|f-g\|$, and condition (3.1) holds for $\mathcal{F}^{(r)}_{L,N}$, with some $M \le 1+\sqrt{(2N+1)L}$. □

It should be stressed that the necessity of imposing the additional restrictive assumption 8 to cure the problems with the dense minimizer of empirical risk does not limit the practical applicability of the main part of the Klemelä-Mammen theory, which is build for δ-net minimizers. The latter not only applies to several typical function classes, as discussed in Section 5 in [9], but also has the additional appeal of being easier to implement. The results on dense minimization of empirical risk, both ours and those in [9], are thus mainly of theoretical interest.

Acknowledgments

The second author was partially supported by the Polish Ministry of Science and Higher Education via an AGH local grant 10.420.03.

The authors thank an anonymous reviewer for suggestions of improvements in presentation of the results.

REFERENCES

[1] L. Birgé, P. Massart, *Rates of convergence for minimum contrast estimators*, Probab. Theory Relat. Fields **97** (1993), 113–150.

[2] L. Birgé, P. Massart, *An adaptive compression algorithm in Besov spaces*, Const. Approx. **16** (2000), 1–36.

[3] A. Cohen, W. Dahmen, I. Daubechies, R. DeVore, *Tree approximation and optimal encoding*, Appl. and Comp. Harmonic Analysis **11** (2001), 192–226.

[4] F. Comte, M.-L. Taupine, Y. Rosenholc, *Penalized contrast estimator for density deconvolution*, Canad. J. Statist. **34** (2006), 431–452.

[5] R. DeVore, G. Lorentz, *Constructive Approximation*, Springer-Verlag, New York, 1993.

[6] E. Gassiat, R. van Handel, *The local geometry of finite mixtures*, Trans. Amer. Math. Soc. **366** (2014), 1047–1072.

[7] J. Heinonen, *Lectures on Analysis on Metric Spaces*, Springer-Verlag, New York, 2001.

[8] J. Klemelä, E. Mammen, *Empirical risk minimization in inverse problems: Extended technical version*, (2009) available at http://arxiv.org/abs/0904.2977v1.

[9] J. Klemelä, E. Mammen, *Empirical risk minimization in inverse problems*, Ann. Statist. **38** (2010), 482–511.

[10] M. Kosorok, *Introduction to Empirical Processes and Semiparametric Inference*, Springer-Verlag, New York, 2008.

[11] L.M. Le Cam, *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York, 1986.

[12] T. Nicoleris, Y.G. Yatracos, *Rates of convergence of estimates, Kolmogorov's entropy and the dimensionality reduction principle in regression*, Ann. Statist. **25** (1997), 2493–2511.

[13] M.J. van der Laan, S. Dudoit, A.W. van der Vaart, *The cross-validated adaptive epsilon-net estimator*, Statist. Decisions **24** (2006), 373–395.

[14] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[15] Y.G. Yatracos, *Rates of convergence of minimum distance estimators and Kolmogorov's entropy*, Ann. Statist. **13** (1985), 768–774.

Jacek Podlewski
podlewski.jacek@gmail.com

StatSoft Poland
ul. Kraszewskiego 36
30-110 Krakow, Poland

Zbigniew Szkutnik
szkutnik@agh.edu.pl

AGH University of Science and Technology
Faculty of Applied Mathematics
al. Mickiewicza 30, 30-059 Krakow, Poland

Received: October 15, 2015.
Revised: March 10, 2016.
Accepted: March 10, 2016.