Electronic Journal of Statistics
Vol. 8 (2014) 543–574
ISSN: 1935-7524
DOI: 10.1214/14-EJS894

On the uniform convergence of empirical norms and inner products, with application to causal inference

Sara van de Geer

Seminar for Statistics
ETH Zürich
e-mail: geer@stat.math.ethz.ch

Abstract: Uniform convergence of empirical norms - empirical measures of squared functions - is a topic which has received considerable attention in the literature on empirical processes. The results are relevant as empirical norms occur due to symmetrization. They also play a prominent role in statistical applications. The contraction inequality has been a main tool but recently other approaches have shown to lead to better results in important cases. We present an overview including the linear (anisotropic) case, and give new results for inner products of functions. Our main application will be the estimation of the parental structure in a directed acyclic graph. As intermediate result we establish convergence of the least squares estimator when the model is wrong.

MSC 2010 subject classifications: Primary 62G08; secondary 60G50.
Keywords and phrases: Additive model, causal inference, empirical measure, uniform convergence.

Received October 2013.

1. Introduction

Let $X_1, \ldots, X_n$ be independent random variables with values in $\mathcal{X}$ and $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. For a function $f : \mathcal{X} \to \mathbb{R}$, we denote its empirical measure by $P_n f := \sum_{i=1}^n f(X_i)/n$ and its theoretical measure by $Pf := \sum_{i=1}^n \mathbb{E} f(X_i)/n$ (assuming it exists). Furthermore, we let $\|f\|_n^2 := P_n f^2$ and $\|f\|^2 := Pf^2$ (again assuming it exists). We call $\|f\|_n$ the empirical norm of the function f and $\|f\|$ its theoretical norm. We review some results concerning the the uniform (over $\mathcal{F}$) convergence of $\|\cdot\|_n$ to $\|\cdot\|$. As example, we consider the case $\mathcal{X} = \mathbb{R}^p$ (with p possibly large) and $\mathcal{F}$ is a class of additive functions $f(x_1, \ldots, x_p) = \sum_{k=1}^p f_0(x_k)$ with f_0 in a given class of functions $\mathcal{F}_0$ on $\mathbb{R}$ (Theorem 2.3). We extend the results to uniform convergence of the empirical measure of products of functions. The latter will be an important tool for statistical theory for causal inference. As intermediate step we show convergence of the least squares estimator when the model is wrong.

In Theorem 2.1 we present results from [12] and [5] and in Theorem 2.2 we compare these with more classical approaches using e.g. the contraction inequality. The extension to inner products is given in Theorem 3.1. The latter

can be used in statistical applications where functions from different smoothness classes are estimated (for example in an additive model).

We pay some special attention to the linear case, i.e. the case where $\mathcal{F}$ is (a subset of) a linear space. For isotropic distributions the uniform convergence of $\|\cdot\|_n$ to $\|\cdot\|$ over linear functions is well developed. We refer to [1] and with sub-Gaussian random vectors to [24, 16] and [26]. We will not require isotropic distributions but instead consider possibly anisotropic but bounded random variables. We present results from [5] and [26] which are based on [12] or a similar approach. Theorems 4.1 and 4.2 are essentially in [5] and [26]. We compare the bound with a Bernstein type inequality for random matrices as given in [3].

Uniform convergence of empirical norms and inner products has numerous statistical applications. This study is motivated by some questions arising in the structural equations model for causal inference. Let us briefly sketch the problem. Consider having observed an $n \times p$ matrix data matrix X with i.i.d. rows. We assume the structural equations model

$$X_{1,j} = f_j^0\Big(\text{parents}(X_{1,j})\Big) + \epsilon_{1,j},\ j = 1, \ldots, p.$$

Here, parents$(X_{1,j})$ is a subset of $\{X_{1,k}\}_{k \neq j}$, $\epsilon_{1,j}, \ldots, \epsilon_{1,p}$ are independent Gaussian noise terms, $\epsilon_{1,j}$ is independent of parents$(X_{1,j})$ and f_j^0 is the regression of $X_{1,j}$ on its parents $(j = 1, \ldots, p)$. For a directed acyclic graph (DAG) there exists a permutation $\pi^0 := (\pi_1^0, \ldots, \pi_p^0)$ of $\{1, \ldots, p\}$ such that for all j the parents of X_{1,π_j^0} are $\{X_{1,\pi_1^0}, \ldots, X_{1,\pi_{j-1}^0},\}$ or a subset thereof, with the convention that for $j = 1$, the parental set is the empty set. The permutation π_0 is not unique, and we let Π_0 be the class of permutations with this parental structure.

If for each j the set of parents of $X_{1,j}$ in the DAG were known, the problem is a standard (nonparametric) multiple regression problem. However, the parental structure, i.e. the class Π_0 is not known and hence has to be estimated from the data. Let Π the class of all $p!$ permutations of $\{1, \ldots, p\}$ and $\{\mathcal{F}_j\}_{j=1}^p$ be given classes of regression functions. Here, $\mathcal{F}_j$ is a collection of functions of $j-1$ variables $(j = 1, \ldots, p)$. We use the short hand notation: for each i, j and π

$$f_j(X_i, \pi) := f_j(X_{i,\pi_1}, \ldots, X_{i,\pi_{j-1}}),$$

with the above convention for $j = 1$, and for each j and π

$$\|\mathbf{X}_{\pi_j} - f_j(\pi)\|_n^2 := \sum_{i=1}^n (X_{i,\pi_j} - f_j(X_i, \pi))^2/n.$$

We consider the estimator

$$\hat{\pi} :\in \arg\min_{\pi \in \Pi} \sum_{j=1}^p \log\Big(\|\mathbf{X}_{\pi_j} - \hat{f}_j(\pi)\|_n\Big)$$

where, for each j, $\hat{f}_j(\pi)$ is the least squares estimator

$$\hat{f}_j(\pi) := \arg\min_{f_j \in \mathcal{F}_j} \|\mathbf{X}_{\pi_j} - f_j(\pi)\|_n.$$

This estimator is proposed by [8], where consistency results, algorithms and simulations are presented. We further develop the theory using the refined inequalities from [12] and [3]. We show in Theorem 6.1 that this estimator is consistent under various scenario's: $\mathbb{P}(\hat{\pi} \notin \Pi_0)$ converges to zero. An important assumption here is an identifiability assumption: see Condition 6.1. This excludes the Gaussian linear structural equations model where $X_{1,j}$ depends linearly on its parents. We will instead model each $f_j \in \mathcal{F}_j$ as being an additive non-linear function

$$f_j(x_1, \ldots, x_{j-1}) = \sum_{k=1}^{j-1} f_{k,j}(x_k),$$

where each $f_{k,j}$ belongs to a given class $\mathcal{F}_0$ of real-valued functions on $\mathbb{R}$.

We consider several cases. The results can be found in Theorem 6.1. They are a consequence of uniform convergence of empirical norms of a class of additive functions as given in Theorem 2.3 which may be of independent interest. Let us summarize the findings here.

In the first two cases, the class $\mathcal{F}_0$ is assumed to have finite entropy integral for the supremum norm. We then derive consistency when $p^3 = o(n)$. Under additional assumptions this is can be relaxed to $p^{3-(1-\alpha)^2} = o(n)$, where $0 < \alpha < 1$ is a measure of the "smoothness" of the class $\mathcal{F}_0$.

An important special case is where $\mathcal{F}_0$ is a class of linear functions. Each $f_{k,j}$ is then a linear combination of functions in a given dictionary $\{\psi_r\}_{r=1}^N$:

$$f_{k,j}(x_k) = \sum_{r=1}^{N} \beta_{r,k,j} \psi_r(x_k).$$

In other words, the dependence of a variable (index j) on one of its parents (index k) is then modelled as a linear combination of certain features (index r) of this parent. We assume the dictionary to be bounded in supremum norm.

If $\mathcal{F}_0$ is the signed convex hull of the functions $\{\psi_r\}_{r=1}^N$ we obtain consistency when $p^2 \log N \log^3 n = o(n)$. The latter situation covers for example the case where $\mathcal{F}_0$ is a collection of functions with total variation bounded by a fixed constant.

Under certain eigenvalue conditions we find that $pN^2 \log n = o(n)$ also yields consistency.

Finally, if $\mathcal{F}_0$ can be approximated by linear functions in a space of dimension N with bias of order $N^{-1/(2\alpha)}$, then consistency follows from $p^{1+4\alpha} \log n = o(n)$.

The paper [8] shows consistency for the case p fixed (the low-dimensional case). It also has theoretical results for the high-dimensional case, but for a restricted estimator where it is assumed that $X_{1,j}$ has only a few parents and a superset of the parents parents$(X_{1,j})$ is known or can be estimated ($j = 1, \ldots, p$). This superset then is required to be small.

The paper is organized as follows. In Sections 2 and 3 we study a generic class of functions $\mathcal{F}$ satisfying some $\|\cdot\|$- and $\|\cdot\|_\infty$-bounds. We present the uniform convergence for empirical norms in Section 2, with main example in Subsection 2.5.

Section 3 looks at empirical inner products of functions in different "smoothness" classes. Subsection 3.2 illustrates the results by considering two classes of functions satisfying different entropy conditions. In many applications one also needs uniform convergence of inner products with a sub-Gaussian (instead of bounded) random variable. Therefore we briefly review this case as well in Subsection 3.3.

Section 4 applies the theory to a class of linear functions and Section 5 studies linear regression when the model is wrong. Section 6 contains the main application: estimation of the order in a directed acyclic graph. Section 7 concludes.

Section 8 presents the technical tools and Section 9 contains the proofs. Throughout C_0, C_1, C_2, ... and c_0, c_1, c_2, ... are universal constants, not the same at each appearance.

2. Bounds for the empirical norm

For a subset S of a metric space (Λ, d) and constant $u > 0$ the u-covering number $N(u, S, d)$ is defined as the minimum number of balls with radius u necessary to cover S, i.e.

$$N(u,S,d) := \min\left\{N : \ \exists\ s_1,\ldots,s_N \text{ such that} \sup_{s\in S}\min_{1\le j\le N} d(s,s_j) \le u\right\}.$$

The entropy of S is $H(\cdot, S, d) := \log N(\cdot, S, d)$. If d is a metric induced by some norm, say τ, we write $N(\cdot, S, \tau)$ $(H(\cdot, S, \tau))$ for the covering numbers (entropy).

2.1. Entropy and entropy integrals

For a real-valued function f on $\mathcal{X}$ we let its supremum norm restricted to the sample be

$$\|f\|_{n,\infty} := \max_{1\le i\le n} |f(X_i)|$$

and we let $\mathcal{H}(u, \mathcal{F}, \|\cdot\|_{n,\infty})$ be the entropy of $(\mathcal{F}, \|\cdot\|_{n,\infty})$. We further define for $z > 0$

$$J_\infty^2(z,\mathcal{F}) := C_0^2 \inf_{\delta>0} \mathbb{E}\left[z\int_\delta^1 \sqrt{\mathcal{H}(uz/2,\mathcal{F},\|\cdot\|_{n,\infty})}du + \sqrt{n}\delta z\right]^2 \tag{2.1}$$

where the constant C_0 is taken as in Theorem 8.3 (Dudley's Theorem). We can without loss of generality assume the integral exists (replace the entropy by a continuous upper bound). The subscript ∞ here refers to the fact that we are considering ℓ_∞-norms.

We also consider uniform ℓ_2-entropies, defined as follows. Let $\mathcal{A}_n$ be the set of all configurations A_n of n (possibly non-distinct) points within the support of P. For $A_n \in \mathcal{A}_n$ and f a real-valued function on $\mathcal{X}$ we let

$$\|f\|_{A_n}^2 := \sum_{x\in A_n} f^2(x)/n.$$

Note that $\|f\|_n = \|f\|_{\mathbf{X}}$ where $\mathbf{X}$ is the random sample $\mathbf{X} := \{X_1, \ldots, X_n\}$. For a class $\mathcal{F}$ of functions on $\mathcal{X}$, we let

$$\mathcal{H}(\cdot, \mathcal{F}) := \sup_{A_n \in \mathcal{A}_n} \mathcal{H}(\cdot, \mathcal{F}, \|\cdot\|_{A_n})$$

and

$$\mathcal{J}_0(z, \mathcal{F}) := C_0 z \int_0^1 \sqrt{\mathcal{H}(uz/2, \mathcal{F})} du, \ z > 0. \tag{2.2}$$

The calligraphic symbol $\mathcal{J}$ indicates that instead of random entropies we consider the maximum entropy over all possible configurations of (at most) n points. Apart from this and from considering ℓ_2-entropy instead of ℓ_∞-entropy we now moreover implicitly assume that the entropy integral converges and use $\mathcal{J}_0$ with subscript 0 to indicate this. The reason for taking 0 as lower-integrant is that $v \mapsto \mathcal{J}_0(\sqrt{v}, \mathcal{F})$ is a concave function. We will see this to be useful in Theorem 2.2 in view of Jensen's inequality.

Finally, for $A_n \in \mathcal{A}_n$ and f a real-valued function on $\mathcal{X}$ we let

$$\|f\|_{A_n,\infty} := \max_{x \in A_n} |f(x)|.$$

Note that $\|f\|_{n,\infty} = \|f\|_{\mathbf{X},\infty}$ where $\mathbf{X}$ is the sample $\mathbf{X} := \{X_1, \ldots, X_n\}$. For a class $\mathcal{F}$ of functions on $\mathcal{X}$ we set

$$\mathcal{H}_\infty(\cdot, \mathcal{F}) := \sup_{A_n \in \mathcal{A}_n} \mathcal{H}(\cdot, \mathcal{F}, \|\cdot\|_{A_n,\infty}).$$

We furthermore define for $z > 0$

$$\mathcal{J}_\infty(z, \mathcal{F}) := C_0 \inf_{\delta > 0} \left[z \int_{\delta/4}^1 \sqrt{\mathcal{H}_\infty(uz/2, \mathcal{F})} du + \sqrt{n}\delta z \right]. \tag{2.3}$$

By the definition of J_∞ (see (2.1)) $J_\infty(z, \mathcal{F}) \leq \mathcal{J}_\infty(z, \mathcal{F})$. We use the calligraphic symbol $\mathcal{J}_\infty$ with subscript ∞ here to indicate that the maximal ℓ_∞-entropy over all possible configurations of (at most) n points is used.

Example 2.1. For real-valued functions g and h with common domain, we use the notation $g \asymp h$ if g/h and h/g are bounded in sup-norm by a constant not depending on n. Let $\mathcal{X} := \mathbb{R}$ and let $\mathcal{F}$ be the class of all increasing functions $f : \mathbb{R} \to [0, 1]$. Then from [7]

$$\mathcal{H}(u, \mathcal{F}) \asymp \frac{1}{u}, \ u > 0,$$

whereas by [31]

$$\mathcal{H}_\infty(u, \mathcal{F}) \asymp \frac{\log n}{u}, \ u > 0.$$

It follows that $\mathcal{J}_0(z) \asymp \sqrt{z}$ and $\mathcal{J}_\infty(z) \asymp \sqrt{z \log n}$. We further note that $\mathcal{F}$ endowed with the metric induced by the supremum norm $\|\cdot\|_\infty$ does not have a finite entropy.

2.2. Bounds using ℓ_∞-norms

The following theorem follows from [12]. Recall the definition (2.1) of J_∞.

Theorem 2.1. *Let*

$$R := \sup_{f\in\mathcal{F}} \|f\|,\ K := \sup_{f\in\mathcal{F}} \|f\|_\infty,\ \hat{R} := \sup_{f\in\mathcal{F}} \|f\|_n.$$

Then

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\|f\|_n^2 - \|f\|^2\right|\right) \le 2J_\infty(K,\mathcal{F})R/\sqrt{n} + 4J_\infty^2(K,\mathcal{F})/n.$$

Moreover, for all $t > 0$, with probability at least $1 - \exp[-t]$,

$$\sup_{f\in\mathcal{F}}\left|\|f\|_n^2 - \|f\|^2\right|/C_1 \le \frac{2RJ_\infty(K,\mathcal{F}) + RK\sqrt{t}}{\sqrt{n}} + \frac{4J_\infty^2(K,\mathcal{F}) + K^2 t}{n}$$

where the constant C_1 is as in Theorem 8.4 (a deviation inequality). As by-product of the proof, we find

$$\sqrt{\mathbb{E}\hat{R}^2} \le R + 2J_\infty(K,\mathcal{F})/\sqrt{n}.$$

Actually, in [12] the entropy integral related quantity J_∞ is replaced by a more general quantity coming from generic chaining.

2.3. Bounds using ℓ_2-norms

In Theorem 2.2 below, we reverse the role of R and K as compared to Theorem 2.1. The result is well-known, it follows from contraction inequality [15] or from a direct argument. See also [11]. Recall the definition (2.2) of $\mathcal{J}_0$. For an increasing convex function $G : [0,\infty) \to [0,\infty)$ with $G(0) = 0$ the convex conjugate H is defined as

$$H(v) := \sup_{u\ge 0}\{uv - G(u)\},\ v \ge 0.$$

Theorem 2.2. *Let*

$$R := \sup_{f\in\mathcal{F}} \|f\|,\ K := \sup_{f\in\mathcal{F}} \|f\|_\infty,\ \hat{R} := \sup_{f\in\mathcal{F}} \|f\|_n.$$

Let for $z > 0$, $G^{-1}(z^2) := \mathcal{J}_0(z,\mathcal{F})$ and let H be the convex conjugate of G. Assume that $R^2 \ge H(4K/\sqrt{n})$. Then

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\|f\|_n^2 - \|f\|^2\right|\right) \le \frac{2K\mathcal{J}_0(2R,\mathcal{F})}{\sqrt{n}}.$$

Moreover, for $R^2 \geq H(4K/\sqrt{n})$ *and all* $t > 0$

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\|f\|_n^2 - \|f\|^2\right|/C_1 \geq \frac{2K\mathcal{J}_0(2R,\mathcal{F}) + KR\sqrt{t}}{\sqrt{n}} + \frac{K^2 t}{n}\right) \leq \exp[-t]$$

where the constant C_1 *is as in Theorem 8.4.*

As by-product of the proof, we find

$$\mathbb{E}\hat{R}^2 \leq 4R^2.$$

2.4. The scaling phenomenon

As said, the essential difference between Theorems 2.1 and 2.2 is that the roles of K and R are reversed, instead of $RJ_\infty(K,\mathcal{F})$ we are dealing with $K\mathcal{J}_0(R,\mathcal{F})$. In some situations $J_\infty(K,\mathcal{F})/K$ behaves as a constant whereas $\mathcal{J}_0(R,\mathcal{F})/R$ decreases in R. Let us illustrate this here. Let $\mathcal{F}_1$ be a class of functions, uniformly $\|\cdot\|_\infty$-bounded by 1, and consider for $R \leq 1$ the localized class

$$\mathcal{F}(R) := \{f \in \mathcal{F}_1 : \ \|f\| \leq R\}.$$

Suppose that

$$\sup_{f\in\mathcal{F}_1(R)} \|f\|_\infty \asymp 1, \ 0 < R \leq 1$$

and for some $0 < \alpha < 1$

$$J_\infty(z,\mathcal{F}_1(R)) \asymp \mathcal{J}_0(z,\mathcal{F}_1(R)) \asymp z^{1-\alpha}, \ z > 0, \ 0 < R \leq 1.$$

These assumptions say that the local class $\mathcal{F}(R)$ behaves like the global class $\mathcal{F}_1$ as far as supremum norm and entropy are concerned. Then, taking $K \asymp 1$,

$$RJ_\infty(K,\mathcal{F}(R)) \asymp R, \ K\mathcal{J}_0(R,\mathcal{F}(R)) \asymp R^{1-\alpha}.$$

Thus, by using Theorem 2.1 instead of Theorem 2.2 we win a factor R^α.

Otherwise put, let $\mathcal{F}_K := \{Kf : \ f \in \mathcal{F}_1\}$ for some $K \geq 1$ and

$$\mathcal{F}_K(1) := \{f \in \mathcal{F}_K : \ \|f\| \leq 1\}.$$

Then, taking $R = 1$,

$$RJ_\infty(K,\mathcal{F}_K(1)) \asymp K, \ K\mathcal{J}_0(R,\mathcal{F}_K(1)) \asymp K^{1+\alpha}.$$

So by using Theorem 2.1 instead of Theorem 2.2 we get rid of a factor K^α.

In fact, we find a scaling phenomenon in Theorem 2.1: whereas for general deviation inequalities the term involving the expectation of the supremum of the empirical process dominates the deviation term, in the current situation they are of the same order.

Also more generally Theorem 2.1 gives better results than Theorem 2.2. As we will see, in the particular case where $\mathcal{F}_1$ is the signed convex hull of p given functions, uniform convergence over $\mathcal{F}_K(1)$ follows from Theorem 2.1 for K of small order $\sqrt{n}$ (up to log-factors) (see Theorem 4.1), whereas Theorem 2.2 needs K to be of small order $n^{1/4}$ (up to log-factors). This corresponds roughly to the above phenomenon with $\alpha \uparrow 1$.

2.5. Example: Additive functions

Let $\mathcal{F}_0$ be a class of real-valued functions defined on the real line. Let further $\mathcal{X} := \mathbb{R}^p$ where $p \leq n$ and let

$$\mathcal{F} := \Big\{ f(x_1, \ldots, x_p) = \sum_{k=1}^{p} f_k(x_k) : \ f_k \in \mathcal{F}_0 \ \forall \ k \Big\}.$$

We will sometimes require the following incoherence condition: for a constant c_1 and for all $f_0 \in \mathcal{F}_0$ and $f_{0,k}(x_1, \ldots, x_p) := f_0(x_k)$, $k = 1, \ldots, p$,

$$\sum_{k=1}^{p} \|f_{0,k}\|^2 \leq c_1 \| \sum_{k=1}^{p} f_{0,k}\|^2. \tag{2.4}$$

For example, if the functions $\{f_{0,k} : k = 1, \ldots, p\}$ can be expanded in terms of orthonormal basis functions $\{\psi_{r,k} : r = 1, \ldots, N, \ k = 1, \ldots, p\}$ this condition is met if the $(Np)\times(Np)$ inner product matrix $(P\psi_{r_1,k_1}\psi_{r_2,k_2})_{\{1\leq r_1,r_2\leq N, \ 1\leq k_1,k_2\leq p\}}$ has smallest eigenvalue $1/c_1$.

In the following theorem one may think of $\mathcal{F}_0$ being for a given $m \in \mathbb{N}$ the Sobolev class

$$\mathcal{F}_0 = \Big\{ f_0 : \ [0,1] \to [0,1] : \ \int_0^1 |f_0^{(m)}(v)|^2 dv \leq 1 \Big\}. \tag{2.5}$$

The constant α is then $\alpha = 1/(2m)$ (see [7]) and the choice $N \asymp n^{\frac{\alpha}{1+\alpha}}$ corresponds to taking a piecewise polynomial approximation with $\asymp n^{\frac{1}{2m+1}}$ pieces (i.e. the bandwidth of the usual order $n^{-\frac{m}{2m+1}}$). The bound (2.7) is shown for this case in [2] under the condition that the X_i have support $[0,1]^p$ and their one-dimensional marginal densities stay away from zero (see also Lemma 2.1 below).

We note that one may also consider variables X_i with unbounded support and classes of functions $\mathcal{F}_0$ on the whole real line. Entropy results for the supremum norm for such cases are in [21].

We define

$$\mathbf{Z}(\mathcal{F}(1)) := \sup_{f\in\mathcal{F}, \ \|f\|\leq 1} \Big| \|f\|_n^2 - \|f\|^2 \Big|.$$

Theorem 2.3.

Case 1. *Assume that for a fixed* $0 < \alpha < 1$,

$$\int_0^z \sqrt{H(u, \mathcal{F}_0, \|\cdot\|_\infty)} du \asymp z^{1-\alpha}, \ z > 0. \tag{2.6}$$

Then $\mathbf{Z}^2(\mathcal{F}(1)) = \mathcal{O}_{\mathbb{P}}(p^3/n)$.

Case 2. *Assume in addition to the condition of Case 1 that the incoherence condition (2.4) holds true for some constant* $c_1 = \mathcal{O}(1)$ *and that for some constant* $c_2 = \mathcal{O}(1)$ *and for all* $f_0 \in \mathcal{F}_0$, *all* j, *and for* $f_{0,k}(x_1, \ldots, x_p) = f_0(x_k)$

$$\|f_{0,k}\|_\infty \leq c_2 \|f_{0,k}\|^{1-\alpha}. \tag{2.7}$$

Then $\mathbf{Z}^2(\mathcal{F}(1)) = \mathcal{O}_{\mathbb{P}}(p^{3-(1-\alpha)^2}/n)$,

Case 3. *Suppose that $\mathcal{F}_0$ is the signed convex hull of given functions $\{\psi_r\}_{r=1}^N$, $N \le n$, in particular*

$$\mathcal{F}_0 = \Big\{ f_0 = \sum_{r=1}^N \beta_r \psi_r(\cdot) : \ \sum_{r=1}^N |\beta_r| \le 1 \Big\}, \tag{2.8}$$

where $\{\psi_r\}$ is a given dictionary satisfying $\max_r \|\psi_r\|_\infty = \mathcal{O}(1)$. Then $\mathbf{Z}^2(\mathcal{F}(1)) = \mathcal{O}_{\mathbb{P}}(p^2 \log^4 n/n)$.

Case 4. *Suppose that for some $N \le n$*

$$\mathcal{F}_0 = \Big\{ f_0(\cdot) = \sum_{r=1}^N \beta_r \psi_r(\cdot) : \ (\beta_1, \dots, \beta_N) \in \mathbb{R}^N \Big\},$$

where $\{\psi_r\}$ is a given dictionary satisfying $\max_r \|\psi_r\|_\infty = \mathcal{O}(1)$. Assume that the incoherence condition (2.4) is met for some constant $c_1 = \mathcal{O}(1)$. Assume moreover that for a constant $c_0 = \mathcal{O}(1)$, all $\beta \in \mathbb{R}^N$, and for all k and for $f_{\beta,k}(x_1, \dots, x_p) := \sum_{r=1}^N \beta_r \psi_r(x_k)$,

$$\|\beta\|_2^2 \le c_0 N \|f_{\beta,k}\|^2. \tag{2.9}$$

Then $\mathbf{Z}^2(\mathcal{F}(1)) = \mathcal{O}_{\mathbb{P}}(pN^2 \log n/n)$. When one chooses $N \asymp n^{\frac{\alpha}{1+\alpha}}$ this reads $\mathbf{Z}^2(\mathcal{F}(1)) = \mathcal{O}_{\mathbb{P}}(pn^{\frac{\alpha}{1+\alpha}} \log n / n^{\frac{1}{1+\alpha}})$.

Case 5. *Consider a dictionary $\{\psi_r\}_{r=1}^\infty$ with $\sup_r \|\psi_r\|_\infty = \mathcal{O}(1)$. Suppose that for a constant $c_0 = \mathcal{O}(1)$ and any $f_0 \in \mathcal{F}_0$ and any $N \in \mathbb{N}$ there exists a $\beta \in \mathbb{R}^N$ such that*

$$\Big\| f_0 - \sum_{r=1}^N \beta_r \psi_r \Big\|_\infty \le c_0 N^{-\frac{1}{2\alpha}}.$$

Moreover, assume the incoherence condition (2.4) with $c_1 = \mathcal{O}(1)$and that for all $N \in \mathbb{N}$, all $\beta \in \mathbb{R}^N$, all k and for $f_{\beta,k}(x_1, \dots, x_p) := \sum_{r=1}^N \beta_r \psi_r(x_k)$,

$$\|\beta\|_2^2 \le c_0 N \|f_{\beta,k}\|^2.$$

Then $\mathbf{Z}^2(\mathcal{F}(1)) = \mathcal{O}_{\mathbb{P}}(p^{\frac{1+4\alpha}{1+2\alpha}} (\log n/n)^{\frac{1}{1+2\alpha}})$.

Remark 2.1. If Condition (2.4) holds, one may in fact replace $\mathcal{F}_0$ in (2.6) of Case 1 by the local class $\mathcal{F}_0 \cap \{\min_k \|f_0\|_k^2 \le c_1\}$ where $\|f_0\|_k^2 = \mathbb{E} \sum_{i=1}^n f_0^2(X_k)/n$:

$$\int_0^z \sqrt{H(u, \mathcal{F}_0 \cap \{\min_k \|f_0\|_k^2 \le c_1\}, \|\cdot\|_\infty)} du \asymp z^{1-\alpha}, \ z > 0.$$

The same is true for Case 2, where Condition (2.4) is indeed assumed. In Case 3, assuming (2.4) one may replace condition (2.8) by the local version

$$\mathcal{F}_0 \cap \{\min_k \|f_0\|_k^2 \le c_2\} \subset \Big\{ f_0 = \sum_{r=1}^N \beta_r \psi_r(\cdot) : \ \sum_{r=1}^N |\beta_r| \le 1 \Big\}.$$

To complete the picture we show in the next lemma that condition (2.7) is natural in the context of Case 5 (although we do not use it there).

Lemma 2.1. *Assume the conditions of Case 5 in Theorem 2.3 and that for some constant $c_3 = \mathcal{O}(1)$, for all r and all $s > c_3$,*

$$\psi_r \psi_{r+s} = 0. \tag{2.10}$$

i.e., that as soon as $s > c_0$, ψ_r and ψ_{r+s} do not overlap. Then (2.7) holds for some constant $c_2 = \mathcal{O}(1)$.

In Case 2, the bound found in [19] is $\mathbf{Z}(\mathcal{F}(1)) = \mathcal{O}_{\mathbb{P}}(p^{2(1+\alpha)}/n)$. Note that in Case 5, we have $\mathbf{Z}(\mathcal{F}(1)) = o_{\mathbb{P}}(1)$ whenever $p^{1+4\alpha}/n = o(1)$. The conditions on p can possible be weakened (possibly by replacing entropy bounds by Gaussian means) but this is an open problem. It is not clear to us whether the bounds presented in Theorem 2.3 are sharp.

Case 1 and 2 of Theorem 2.3 follow from Theorem 2.1 by straightforward entropy bounds. Case 3 is based on a result from [26] cited here as Theorem 4.1. Case 4 is based on the general matrix version of Bernstein's inequality of [3] cited here as Theorem 4.3. Case 5 follows from Case 4 using a trade-off argument for the choice of N (the value $N = n^{\alpha/(1+\alpha)}$ suggested in Case 4 may not give the optimal trade-off). The details are in Section 9.

3. Empirical inner products

Consider products fg of functions f and g with f in some class $\mathcal{F}$ and g in some class $\mathcal{G}$. Note that one can derive results for products via squares:

$$fg = (f+g)^2/2 - (f^2 + g^2)/2.$$

If $\mathcal{F}$ and $\mathcal{G}$ have the same $\|\cdot\|$-diameter R and the same $\|\cdot\|_\infty$-diameter K it is easy to see that without loss of generality we may assume that $\mathcal{F} = \mathcal{G}$ (replace $\mathcal{F}$ and $\mathcal{G}$ by $\mathcal{F} \cup \mathcal{G}$). However, if f and g are in different classes it may be more appropriate to analyze the products directly. This case with $\mathcal{F}$ and $\mathcal{G}$ having different radii is studied here.

We only present the results using ℓ_∞-norms. Again, one may reverse the roles of $\|\cdot\|_\infty$-radii and $\|\cdot\|$-radii, getting other versions for the bounds. The best bound may depend on the situation at hand.

3.1. Inner products of functions from different classes

Let

$$R_1 := \sup_{f \in \mathcal{F}} \|f\|, \ K_1 := \sup_{f \in \mathcal{F}} \|f\|_\infty.$$

and

$$R_2 := \sup_{g \in \mathcal{G}} \|g\|, \ K_2 := \sup_{g \in \mathcal{G}} \|g\|_\infty.$$

Theorem 3.1. *Suppose that $R_1K_2 \le R_2K_1$. Consider values of $t \ge 4$ and n such that*

$$\left(\frac{2R_1\mathcal{J}_\infty(K_1,\mathcal{F})+R_1K_1\sqrt{t}}{\sqrt{n}}+\frac{4\mathcal{J}_\infty^2(K_1,\mathcal{F})+K_1^2t}{n}\right)\le\frac{R_1^2}{C_1} \tag{3.1}$$

and

$$\left(\frac{2R_2\mathcal{J}_\infty(K_2,\mathcal{G})+R_2K_2\sqrt{t}}{\sqrt{n}}+\frac{4\mathcal{J}_\infty^2(K_2,\mathcal{G})+K_2^2t}{n}\right)\le\frac{R_2^2}{C_1}. \tag{3.2}$$

Then with probability at least $1-12\exp[-t]$

$$\frac{1}{8C_1}\sup_{f\in\mathcal{F},\ g\in\mathcal{G}}\left|(P_n-P)fg\right|\le\frac{R_1\mathcal{J}_\infty(K_2,\mathcal{G})+R_2\mathcal{J}_\infty(R_1K_2/R_2,\mathcal{F})+R_1K_2\sqrt{t}}{\sqrt{n}}+\frac{K_1K_2t}{n}.$$

Remark 3.1. Theorem 3.1 can be refined using generic chaining type of quantities instead of entropies. We have omitted this to avoid digressions.

Remark 3.2. Consider the special case where $\mathcal{G}=\{g_0\}$ is a singleton. Assume that $\|g_0\|_\infty = K_0$. Take $R_2 = K_2 = K_0$ in Theorem 3.1, and write $R_1 := K$ and $K_1 := K$. For a singleton $\mathcal{G}$, the term $\mathcal{J}_\infty(K_2,\mathcal{G})$ can be omitted. We then get from Theorem 3.1: for $t \ge 4$ and

$$\left(\frac{2R\mathcal{J}_\infty(K,\mathcal{F})+RK\sqrt{t}}{\sqrt{n}}+\frac{4\mathcal{J}_\infty^2(K,\mathcal{F})+K^2t}{n}\right)\le\frac{R^2}{C_1},$$

it holds that

$$\frac{1}{8C_1}\sup_{f\in\mathcal{F}}\left|(P_n-P)fg_0\right|\le\frac{K_0\mathcal{J}_\infty(R,\mathcal{F})+K_0R\sqrt{t}}{\sqrt{n}}+\frac{KK_0t}{n}$$

with probability at least $1-8\exp[-t]$. We will see a similar result in Theorem 3.2, where g_0 is not bounded but sub-Gaussian.

3.2. Empirical inner products for smooth functions

Let us suppose that

$$\mathcal{J}_\infty(z,\mathcal{F})\asymp z^{1-\alpha},\ \mathcal{J}_\infty(z,\mathcal{G})\asymp z^{1-\beta},$$

where $\beta > \alpha$. For example, one may think of Sobolev classes as was indicated in Subsection 2.5, or more locally adaptive cases such as $\mathcal{F}=\{f:[0,1]\to[0,1],\ \int|f''(x)|dx\le 1\}$ and $\mathcal{G}\subset\{g:[0,1]\to[0,1]:\ \int|g'(x)|dx\le 1\}$. Then $\mathcal{J}_\infty(z,\mathcal{F})\asymp z^{3/4}$ ($\alpha=1/4$) and $\mathcal{J}_\infty(z,\mathcal{G})\asymp z^{1/2}\sqrt{\log n}$ ($\beta=1/2$). The $\log n$-term plays a moderate role and we neglect such details in the following general line of reasoning.

The fact that $\beta > \alpha$ expresses that $\mathcal{F}$ is smoother (less rich) than $\mathcal{G}$. Having an additive model in mind (the response Y_i is an additive function plus noise $Y_i = f^0(X_{i,1}) + g^0(X_{i,2}) + \varepsilon_i$, $i = 1, \ldots, n$) one may expect to be able to estimate a function $f^0 \in \mathcal{F}$ with squared rate $R_1^2 := n^{-1/(1+\alpha)}$ and a function $g^0 \in \mathcal{G}$ with (slower) squared rate $R_2^2 := n^{-1/(1+\beta)}$. Let us simplify the situation by assuming that $X_{i,1}$ and $X_{i,2}$ are independent (the dependent case is detailed in [33]). Also assume that the functions in $\mathcal{F}$ and $\mathcal{G}$ are already centred. We now want to show that $P_n fg$ is small, namely negligible as compare to R_1^2. Indeed, inserting Theorem 3.1 (note that (3.1) and (3.2) are true for t fixed and n sufficiently large), we get with probability at least $1 - 12\exp[-t]$

$$\sup_{f \in \mathcal{F}, \|f\| \le R_1,\ g \in \mathcal{G},\ \|g\| \le R_2} \frac{|P_n fg|/c_1}{R_1^2} \le \left(\frac{1}{\sqrt{n} R_1} + R_2^{\alpha} + \sqrt{\frac{t}{n R_1^2}} + \frac{t}{n R_1^2} \right).$$

For fixed t the right hand side of the above inequality is $o(1)$.

Actually, [33] first prove the global (slow) rate $R = R_2$. Suppose that that now $f/K_1 \in \mathcal{F}$ where $K_1 = R/\lambda$ with $\lambda \asymp n^{-1/(1+\alpha)}$. Again (3.1) and (3.2) are true for t fixed and n sufficiently large for $R_1^2 = R_2^2 = R^2 = n^{-1/(1+\beta)}$, $K_1 = R/\lambda$ and $K_2 = 1$. We find as similar result as above: with probability at least $1 - 12\exp[-t]$

$$\sup_{f/K_1 \in \mathcal{F},\ \|f\| \le R,\ g \in \mathcal{G},\ \|g\| \le R} \frac{|P_n fg|/c_1}{R^2} \le \frac{1}{\sqrt{n} R} + \sqrt{\frac{t}{n R^2}} + \frac{t}{n \lambda R}.$$

Related is the paper [20] where the additive model is studied with f^0 a high-dimensional linear function. Again, it can be shown that f^0 can be estimated with a fast oracle rate, faster than the rate of estimation of the unknown function g^0.

3.3. Products with a sub-Gaussian random variable

Consider now real valued random variables Y_i, $i = 1, \ldots, n$. We let P_n be the empirical measure based on $\{X_i, Y_i\}_{i=1}^n$: for a real-valued function f on $\mathcal{X}$

$$P_n \mathbf{Y} f := \sum_{i=1}^n Y_i f(X_i)/n.$$

We write $P\mathbf{Y} f := \mathbb{E} P_n \mathbf{Y} f$. We study the supremum of the absolute value of the product process $(P_n - P)\mathbf{Y} f$, $f \in \mathcal{F}$.

Definition 3.1. For $Z \in \mathbb{R}$ and $\Psi_k(z) := \exp[|z|^k]$, $k = 1, 2$, we define the Orlicz norm

$$\|Z\|_{\Psi_k} := \inf\{L > 0 : \mathbb{E}\Psi_k(Z/L) - 1 < 1\},$$

whenever it exists. If $\|Z\|_{\Psi_1}$ exists, we call Z sub-exponential, and if $\|Z\|_{\Psi_2}$ exists we call Z sub-Gaussian.

Definition 3.2. We say that $\mathbf{Y} := \{Y_1, \ldots, Y_n\}$ is uniformly sub-Gaussian with constant K_0 if

$$\max_{1 \le i \le n} \|Y_i\|_{\Psi_2} \le K_0 < \infty.$$

The result below is about products of functions, where the class $\mathcal{G}$ consists of the single sub-Gaussian function $\mathbf{Y}$.

We recall the definition (2.3) of $\mathcal{J}_\infty$.

Theorem 3.2. *Let*

$$\sup_{f \in \mathcal{F}} \|f\| \le R, \ K := \sup_{f \in \mathcal{F}} \|f\|_\infty.$$

Suppose $\mathbf{Y}$ *is uniformly sub-Gaussian with constant* K_0. *Consider values of* t *and* n *such that*

$$\sqrt{\frac{2t}{n}} + \frac{t}{n} \le 1.$$

For these values

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |(P_n - P)\mathbf{Y} f|/C_1 \ge \frac{2\mathcal{J}_0(KK_0, \mathcal{F}) + KK_0\sqrt{t}}{\sqrt{n}} \right) \le 8 \exp[-t]$$

where the constant C_1 *is as in Theorem 8.5.*

4. Application to a class of linear functions

Suppose $\mathcal{X} = \mathbb{R}^p$. We let X_i be a row vector in $\mathbb{R}^p$, $i = 1, \ldots, n$. For a column vector $\beta \in \mathbb{R}^p$ we define $f_\beta(X_i) := X_i\beta$. We assume in that for some constant K_X

$$\max_{i,j} |X_{i,j}| \le K_X.$$

The following lemma is Lemma 3.7 in [25]. We inserted an explicit constant.

Lemma 4.1. *We have*

$$\mathcal{H}(u, \{f_\beta : \ \|\beta\|_1 \le 1\}, \|\cdot\|_{n,\infty}) \le \left(1 + \frac{8 \log(2p) \log(2n) K_X^2}{u^2}\right), \ u > 0.$$

As a consequence, we obtain a result which is in [26]. It suffices to combine Theorem 2.1 with Lemma 4.1.

Theorem 4.1. *For all* $t > 0$

$$\mathbb{P}\left(\sup_{\|\beta\|_1 \le M, \ \|f_\beta\| \le 1} \left| \|f_\beta\|_n^2 - \|f_\beta\|^2 \right| / c_1 \ge MK_X \sqrt{\frac{\log p \log^3 n + t}{n}} \right.$$

$$\left. + M^2 K_X^2 \frac{\log p \log^3 n + t}{n} \right) \le \exp[-t].$$

Theorem 4.1 has very useful applications, in particular to ℓ_1-regularization or to exact recovery using basis pursuit [9] where results often rely on bounds for compatibility constants [29, 30] or restricted eigenvalues [6]. This is elaborated upon in [26].

Theorem 4.1 can be applied also to obtain a uniform bound over all subspaces. Define the minimal eigenvalue $\Lambda_{\min}^2 := \min_{\|\beta\|_2 \leq 1} \|f_\beta\|^2$.

Theorem 4.2. *Suppose* $\Lambda_{\min} > 0$. *Define for* $S \subset \{1, \ldots, p\}$, $\beta_{j,S} = \beta_j 1\{j \in S\}$, $j = 1, \ldots, p$. *For all* $t > 0$

$$\mathbb{P}\Bigg(\exists\ s: \sup_{|S|=s} \sup_{\|f_{\beta_S}\| \leq 1} \Big| \|f_{\beta_S}\|_n^2 - \|f_{\beta_S}\|^2 \Big| / c_1 \geq \frac{K_X}{\Lambda_{\min}} \sqrt{\frac{s \log p \log^3 n + st}{n}} \tag{4.1}$$

$$+ \frac{K_X^2}{\Lambda_{\min}^2} \frac{s \log p \log^3 n + st}{n} \Bigg) \leq \exp[-t].$$

The next theorem is a direct application of a Bernstein type inequality for random matrices as given in [3] (see also Theorem 3 in [13]). It shows that in Theorem 4.2 the $\log^3 n$-term can be omitted when one considers a fixed set S instead of requiring a result uniform in S.

Theorem 4.3. *Suppose* $\Lambda_{\min} > 0$. *For all* $t > 0$

$$\mathbb{P}\Bigg(\sup_{\|f_\beta\| \leq 1} \Big| \|f_\beta\|_n^2 - \|f_\beta\|^2 \Big| / c_1 \geq \frac{K_X}{\Lambda_{\min}} \sqrt{\frac{p \log p + pt}{n}} \tag{4.2}$$

$$+ \frac{K_X^2}{\Lambda_{\min}^2} \frac{p \log p + pt}{n} \Bigg) \leq \exp[-t].$$

Remark 4.1. Let us briefly indicate how this compares to an isotropic case. Following an idea of [16] (see also Lemma 1 in [22]) one can show that the supremum over all $\|f_\beta\| \leq 1$ can in fact be replaced by a maximum over a finite class:

$$\sup_{\|f_\beta\| \leq 1} \Big| \|f_\beta\|_n^2 - \|f_\beta\|^2 \Big| \leq c_1 \max_{j \in \{1, \ldots, N\}} \Big| (P_n - P) f_{\beta_j}^2 \Big|,$$

where $\|f_{\beta_j}\| \leq 1$ for all $j = 1, \ldots, N$ and where $\log N \leq c_0^2 p$. We can now proceed by invoking the union bound for the maximum. An isotropy assumption then leads to good results. We assume sub-Gaussianity of the vectors $\{X_i\}$, meaning that each $f_\beta(X_i)$ is sub-Gaussian: there is a constant K_1 such that for all $\|f_\beta\| \leq 1$ and all i it holds that $\|f_\beta(X_i)\|_{\Psi_2} \leq K_1$. Then by Bernstein's inequality, for all all $\|f_\beta\| \leq 1$ and all $t > 0$

$$\mathbb{P}\Bigg(\Big| \|f_\beta\|_n^2 - \|f_\beta\|^2 \Big| / C_1 \geq K_1^2 \sqrt{t/n} + K_1^2 t/n \Bigg) \leq \exp[-t].$$

The union bound together with the above reduction then gives for all $t > 0$

$$\mathbb{P}\Big(\sup_{\|f_\beta\|\leq 1}\Big|\|f_\beta\|_n^2 - \|f_\beta\|^2\Big|/(c_1C_1) \geq K_1^2\sqrt{\frac{c_0^2p+t}{n}} + K_1^2\frac{c_0^2p+t}{n}\Big) \leq \exp[-t].$$

The latter result is a "true" deviation inequality: the deviation from the bound $\asymp \sqrt{p/n}$ for the mean does not involve this bound, i.e., there is no p in front of t inside the probability. This in contrast to the result (4.1) in Theorem 4.3.

Remark 4.2. One may wonder why the minimal eigenvalue is playing a role in the result of Theorems 4.2 and 4.3. Of course, as far as conditions on $L_2(P)$-norms are concerned one may orthogonalize the variables. However, after orthogonalization the sup-norm of the variables can be quite large.

The following lemma improves Theorem 3.2 in the linear case.

Lemma 4.2. *Suppose that* $\mathbf{Y} := \{Y_1, \ldots, Y_n\}$ *is uniformly sub-Gaussian with constant* K_0 *(see Definition 3.2). Then for all* $t > 0$

$$\mathbb{P}\Big(\sup_{\|f_\beta\|_1\leq 1}|(P_n - P)\mathbf{Y}f_\beta|/c_2 \geq \frac{K_0K_X}{\Lambda_{\min}}\sqrt{\frac{pt}{n}}\Big) \leq 2\exp[-t].$$

To avoid too involved expressions, we from now on will use order symbols. Then, the results needed for the next section can be summarized as follows.

Summary 4.1. Suppose that $\mathbf{Y} := \{Y_1, \ldots, Y_n\}$ is uniformly sub-Gaussian with constant K_0, that $\max_{i,j}|X_{i,j}| \leq K_X$, $\Lambda_{\min} > 0$ and that $\delta_n = o(1)$, where

$$\delta_n^2 := \frac{K_X^2(1+K_0^2)p\log p}{n\Lambda_{\min}^2}.$$

Then uniformly in $\|f_\beta\| \leq 1$, $\|f_{\tilde{\beta}}\| \leq 1$ is holds that

$$\Big|\|f_\beta\|_n^2 - \|f_\beta\|^2\Big| = \mathcal{O}_{\mathbb{P}}(\delta_n), \ \Big|(P_n - P)(\mathbf{Y} - f_{\tilde{\beta}})f_\beta\Big| = \mathcal{O}_{\mathbb{P}}(\delta_n).$$

5. Least squares when the model is wrong

In this section we examine a p-dimensional linear model with p moderately large, and the least squares estimator. The observations are $\{(X_i, Y_i)\}_{i=1}^n$, independent, and with $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$ $(i = 1, \ldots, n)$. Let $\{\psi_j\}_{j=1}^p$ be a given dictionary of functions on $\mathcal{X}$. We write $f_\beta(\cdot) := \sum_{j=1}^p \beta_j\psi_j(\cdot)$, $\beta \in \mathbb{R}^p$.

The least squares estimator is

$$\hat{f} := \arg\min_{f_\beta}\sum_{i=1}^n(Y_i - f_\beta(X_i))^2.$$

Let $f^0(X_i) := \mathbb{E}(Y_i|X_i)$ be the conditional expectation of Y_i given X_i, $i = 1,\ldots,n$. The projection in $L_2(P)$ of f^0 on the linear space $\{f_\beta : \beta \in \mathbb{R}^p\}$ is written as f^*. We want to show convergence of $\hat{f}$ to f^*. In other words, we aim at convergence to the best approximation of f^0 in the class $\mathcal{F} := \{f_\beta : \beta \in \mathbb{R}^p\}$. This is the statistical learning setup, as in e.g. [17] or [4]. They consider very general classes $\mathcal{F}$ and loss functions but arrive at slower rates or require $\mathcal{F}$ to consist of functions bounded by 1. A related recent work is [14], but also here bounds on the sup-norm in terms of the L_2-norm are imposed. In our setup however, we know little about the higher order moments of f^* (only the second moment is under control as $\|f^*\| \leq \|\mathbf{Y}\|$). The situation is therefore a little more delicate than in the usual regression context (where $\|f^* - f^0\|$ is small). This is where uniform convergence of $\|\cdot\|_n$ to $\|\cdot\|$ comes in.

Lemma 5.1. *Let $0 < \delta_n < 1/2$. On the set*

$$\mathcal{T} := \Big\{ \sup_{\|f_\beta\|\leq 1} \Big| \|f_\beta\|_n^2 - \|f_\beta\|^2 \Big| \leq \delta_n, \sup_{\|f_\beta\|\leq 1,\ \|f_{\tilde{\beta}}\|\leq 1} \Big| 2(P_n - P)(\mathbf{Y} - f_{\tilde{\beta}}) f_\beta \Big| \leq \delta_n \Big\},$$

it holds that

$$\|\hat{f} - f^*\| \leq 2\delta_n.$$

To handle the set $\mathcal{T}$ given in the above lemma, we invoke Summary 4.1. To this end, define the matrix $\Sigma := P\psi^T\psi$ and let $\Lambda_{\min}^2$ be the smallest eigenvalue of Σ.

Theorem 5.1. *Suppose that $\mathbf{Y} := \{Y_1,\ldots,Y_n\}$ is uniformly sub-Gaussian with constant K_0 (see Definition 3.2), that $\max_{i,j} |X_{i,j}| \leq K_X$, $\Lambda_{\min} > 0$, and that $\delta_n = o(1)$ where*

$$\delta_n^2 := \frac{K_X^2(1 + K_0^2)p \log p}{n\Lambda_{\min}^2}.$$

Then

$$\|\hat{f} - f^*\| = \mathcal{O}_{\mathbb{P}}(\delta_n).$$

Moreover

$$\Big| \|\mathbf{Y} - \hat{f}\|_n^2 - \|\mathbf{Y} - f^*\|^2 \Big| = \mathcal{O}_{\mathbb{P}}(\delta_n).$$

In view of the uniformity in Summary 4.1 we can formulate an extension. Such an extension will be useful in the next section. Recall the notation: for a set $S \subset \{1,\ldots,p\}$ and $\beta \in \mathbb{R}^p$

$$\beta_{j,S} := \beta_j\{j \in S\},\ j = 1,\ldots,p.$$

Consider, for any set $S \subset \{1,\ldots,p\}$, the projection f_S^* of f^0 on the $|S|$-dimensional space $\mathcal{F}_S := \{f_{\beta_S} : \beta \in \mathbb{R}^p\}$ and the corresponding least squares estimator

$$\hat{f}_S := \arg\min_{f_{\beta_S}} \|\mathbf{Y} - f_{\beta_S}\|_n.$$

Theorem 5.2. *Assume the conditions of Theorem 5.1 and let δ_n be defined as there. Then uniformly in all S,*

$$\|\hat{f}_S - f_S^*\| = \mathcal{O}_{\mathbb{P}}(\delta_n).$$

Moreover

$$\left| \|\mathbf{Y} - \hat{f}_S\|_n^2 - \|\mathbf{Y} - f_S^*\|^2 \right| = \mathcal{O}_{\mathbb{P}}(\delta_n).$$

6. Application to DAG's

Let X be a $n \times p$ matrix with i.i.d. rows. We throughout this section assume $p \leq n$. The i-th row is denoted by $X_i := (X_{i,1}, \ldots, X_{i,p})$ $(i = 1, \ldots, n)$. The distribution of a row, say X_1, is denoted by P.

We assume a directed acyclic graph (DAG) structure. Namely, we assume the structural equations model defined as follows.

Definition 6.1. We say that $X_1 \in \mathbb{R}^p$ satisfies the non-linear Gaussian structural equations model if for some permutation π^0 of $\{1, \ldots, p\}$ and for some functions $f_j^0 : \mathbb{R}^{j-1} \to \mathbb{R}$

$$X_{1,\pi_j^0} = f_j^0(X_{1,\pi_1^0}, \ldots X_{1,\pi_{j-1}^0}) + \varepsilon_{1,\pi_j^0}, \ j = 1, \ldots, p,$$

where $\{\varepsilon_{1,1}, \ldots, \varepsilon_{1,p}\}$ are independent and where for $j = 1, \ldots, p$ the random variable $\varepsilon_{1,\pi_j^0} \sim \mathcal{N}(0, \sigma^2_{\pi_j^0})$ is independent of $(X_{1,\pi_1^0}, \ldots, X_{1,\pi_{j-1}^0})$. The latter set is to be understood as the empty set when $j = 1$.

We let Π_0 be the set of permutations π_0 for which Definition 6.1 holds. The case of interest is the one where $f_j^0(X_{1,\pi_1^0}, \ldots, X_{1,\pi_{j-1}}) = f_j^*(\{X_{1,k}\}_{k \neq j}\})$, $j = 1, \ldots, p$, with with $\{f_j^*\}$ and hence and $\{\sigma_j^2\}$ not depending on π_0. Our aim is to find a member from Π^0 based on the data $X_1, \ldots, X_n$.

6.1. Some notation

We consider a given class $\mathcal{F}_0$ of functions $f_0 : \mathbb{R} \to \mathbb{R}$.

Let $\mathcal{F}_1 := \emptyset$ and for $j = 2, \ldots, p$

$$\mathcal{F}_j := \Big\{ f(x_1, \ldots, x_{j-1}) = \sum_{k=1}^{j-1} f_{k,j}(x_k) : (x_1, \ldots, x_{j-1}) \in \mathbb{R}^{j-1}, \ f_{k,j} \in \mathcal{F}_0 \ \forall \ j, k \Big\}.$$

Let Π be the set of all permutations of $\{1, \ldots, p\}$. Write for each permutation $\pi \in \Pi$ and each j, for $f_j \in \mathcal{F}_j$,

$$f_j(X_i, \pi) := f_j(X_{i,\pi_1}, \ldots, X_{i,\pi_{j-1}}), \ i = 1, \ldots, n.$$

Define

$$f_j^*(\pi) := \arg\min\left\{ \|\mathbf{X}_{\pi_j} - f_j(\pi)\|^2 : \ f_j \in \mathcal{F}_j \right\}.$$

where for $j = 1$, we take $f_1^*(\pi) = 0$, and where

$$\|\mathbf{X}_{\pi_j} - f_j(\pi)\|^2 := \mathbb{E}\left(X_{1,\pi_j} - f_j(X_1, \pi) \right)^2.$$

We further define

$$\sigma_j^2(\pi) := \|\mathbf{X}_{\pi_j} - f_j^*(\pi)\|^2, \ j = 1, \ldots, p.$$

6.2. Identifiability

In order to be able to estimate a correct permutation one needs to assume that the wrong permutations can be detected.

Condition 6.1 (Identifiability condition)**.** For some constant $\xi > 0$,

$$\inf_{\pi \notin \Pi_0, \ \pi_0 \in \Pi_0} \frac{1}{p} \sum_{j=1}^{p} \log\left(\frac{\sigma_j(\pi)}{\sigma_j(\pi^0)} \right) > \xi.$$

This condition is discussed in [8]. The linear Gaussian structural equations model has $\Pi_0 = \Pi$, i.e. any permutation is correct. In the non-linear case, we think of the situation where, unlike the linear case, the parental dependence is the same for all $\pi_0 \in \Pi_0$, say $f_j^0(X_{1,\pi_1^0}, \ldots, X_{1,\pi_{j-1}^0}) := f_j^*(\{X_{1,k}\}_{k \neq j})$ $(j = 1, \ldots, p)$, and hence also the residual variances σ_j^2, $j = 1, \ldots, p$ do not depend on π_0. The identifiably condition then requires that choosing $\pi \notin \Pi_0$ will give on average too large residual variances. If the model is misspecified, Condition 6.1 is to be seen as assuming robustness to the bias that misspecification introduces. In an asymptotic formulation, it suffices to assume identifiability at the truth: $\inf_{\pi \notin \Pi_0} \sum_{j=1}^{p} \log(\sigma_j(\pi)/\sigma_j)/p > \xi_0$ with $1/\xi_0 = \mathcal{O}(1)$ together with a vanishing bias: $\sup_{\pi_0 \in \Pi_0} \sum_{j=1}^{p} \log(\sigma_j(\pi_0)/\sigma_j)/p \to 0$. One may consider choosing a model with low complexity (large bias) because π_0 is the parameter of interest here. The estimation of f_j^0 $(j = 1, \ldots, p)$ can then follow in a second step using a standard (nonparametric) regression estimator and the estimated permutation.

6.3. The estimator

To describe the estimator of π^0 we introduce empirical counterparts of the quantities given above. For each j and π we write

$$\|\mathbf{X}_{\pi_j} - f_j(\pi)\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} \left(X_{i,\pi_j} - f_j(X_i, \pi) \right)^2.$$

We let $\hat{f}_j(\pi)$ be the least squares estimator

$$\hat{f}_j(\pi) := \arg\min\left\{ \|\mathbf{X}_{\pi_j} - f_j(\pi)\|_n : \ f_j \in \mathcal{F}_j \right\}$$

and take the normalized residual sum of squares

$$\hat{\sigma}_j(\pi) := \|\mathbf{X}_{\pi_j} - \hat{f}_j(\pi)\|_n^2$$

as estimator of $\sigma_j^2(\pi)$. We then let

$$\hat{\pi} :\in \arg\min_{\pi\in\Pi} \sum_{j=1}^{p} \log \hat{\sigma}_j^2(\pi). \tag{6.1}$$

6.4. Consistency

Let $H(\cdot, \mathcal{F}_0, \|\cdot\|_\infty)$ be the entropy of $\mathcal{F}_0$ endowed with supremum norm.

Theorem 6.1. *Suppose the non-linear Gaussian structural equations model (see Definition 6.1) with* $\max_{1\le j\le p}\sigma_j^2 = \mathcal{O}(1)$ *and* $\max_{1\le j\le p}\|f_j^0\|_\infty = \mathcal{O}(1)$. *Assume Condition 6.1 (the identifiability condition) with* $1/\xi = \mathcal{O}(1)$. *Assume moreover that* $\mathcal{F}_0$ *is a convex class and that one of the following 5 cases hold of Theorem 2.3 for the collection* $\mathcal{F} := \{f = \sum_{k=1}^p f_k(x_k),\ f_k \in \mathcal{F}_0\ \forall k\}$*:*

Case 1. *Case 1 holds and* $p^3/n = o(1)$,
Case 2. *Case 2 holds and* $p^{3-(1-\alpha)^2}/n = o(1)$,
Case 3. *Case 3 holds and* $p^2\log^4 n/n = o(1)$,
Case 4. *Case 4 holds and* $pN^2\log n/n = o(1)$,
Case 5. *Case 5 holds,* $p^{1+4\alpha}\log n/n = o(1)$.

Then $\mathbb{P}(\hat{\pi} \notin \Pi_0) \to 0$.

We recall Remark 2.1: the conditions on $\mathcal{F}$ may be weakened to local versions.

7. Conclusion

In this paper we summarized some results for the uniform convergence of empirical norms and the extension to empirical inner products.

For statistical theory the results are very useful. In [5] one can find an application to ℓ_1-restricted regression for the case of random design and [26] focuses on the restricted isometry property and restricted eigenvalues. We have given the application to order estimation in directed acyclic graphs (DAG's). We omitted important computational issues and further discussions for this special case as it is beyond the scope of the paper. For more details we refer to [8].

The results can also be applied to generalize the results in [32] for DAG's to the linear non-Gaussian case, in particular to anisotropic distributions. A generalization to to isotropic distributions (e.g. sub-Gaussian distributions) is possible but perhaps less relevant as in many statistical applications isotropy is not very natural or stable (for DAG's sub-Gaussianity can hold when the linear model is exactly true but it is not clear what happens when the model is only approximately linear).

A further application is the estimation of a precision matrix for non-Gaussian data. We mention that such an approach is used in [34] to construct confidence

intervals for a single parameter. Here, a Lasso is used for estimating a Fisher-information matrix. The estimator is based on empirical projections and also the function to be estimated is a theoretical projection as in Section 5. In the context of confidence intervals in ℓ_2, the uniform convergence may generalize the (sub-)Gaussian case considered in [22]. Another application of uniform convergence, this time for additive models [20, 33], was briefly indicated in Subsection 3.2.

8. Technical tools

8.1. Symmetrization

Define

$$\mathbf{Z}(\mathcal{F}) := \sup_{f\in\mathcal{F}} \Big| (P_n - P)f \Big|.$$

Let moreover $\epsilon_1, \ldots, \epsilon_n$ be a Rademacher sequence (that is, $\epsilon_1, \ldots, \epsilon_n$ are independent random variables taking the values $+1$ or -1 each with probability $\frac{1}{2}$) independent of $X_1, \ldots, X_n$, and define

$$\mathbf{Z}^\epsilon(\mathcal{F}) := \sum_{i=1}^n \epsilon_i f(X_i)/n.$$

Theorem 8.1 (see e.g. [35]). *It holds that*

$$\mathbb{E}\mathbf{Z}(\mathcal{F}) \le 2\mathbb{E}\mathbf{Z}^\epsilon(\mathcal{F}).$$

Theorem 8.2 (see [23]). *Let $R := \sup_{f\in\mathcal{F}} \|f\|$. For $t \ge 4$,*

$$\mathbb{P}(\mathbf{Z}(\mathcal{F}) \ge 4R\sqrt{2t/n}) \le 4\mathbb{P}(\mathbf{Z}^\varepsilon(\mathcal{F}) \ge R\sqrt{2t/n}).$$

8.2. Dudley's theorem

Dudley's theorem is originally for Gaussian processes (see [10]). The extension to sub-Gaussian random variables and Rademacher averages is rather straightforward. We summarize these in our context in Theorem 8.3 below.

Let $\mathcal{H}(\cdot, \mathcal{F}, \|\cdot\|_n)$ denote the entropy of $\mathcal{F}$ equipped with the metric induced by the empirical norm $\|\cdot\|_n$. and let $\hat{R}$ be the random radius $\hat{R} := \sup_{f\in\mathcal{F}} \|f\|_n$.

The theorem below can be found in [15],Theorem 11.1, although not for the case of a diverging entropy integral. This extension is however easily incorporated (see also [28], Lemma 3.2, for a probability inequality with possibly diverging entropy integral).

Theorem 8.3 (Rademacher averages). *We have*

$$\mathbb{E}\mathbf{Z}^\epsilon(\mathcal{F}) \le C_0 \inf_{\delta>0} \mathbb{E}\left[\hat{R}\int_\delta^1 \sqrt{\mathcal{H}(u\hat{R}, \mathcal{F}, \|\cdot\|_n)}du/\sqrt{n} + \delta\hat{R}\right].$$

8.3. Deviation inequalities

We present two deviation inequalities, for the bounded case and the sub-Gaussian case.

Theorem 8.4 (see [27, 18]). *Suppose that for some constants R and K.*

$$\sup_{f\in\mathcal{F}}\|f\|\le R,\ \sup_{f\in\mathcal{F}}\|f\|_\infty\le K.$$

Then for all $t>0$

$$\mathbb{P}\left(\mathbf{Z}(\mathcal{F})/C_1\ge\mathbb{E}\mathbf{Z}(\mathcal{F})+R\sqrt{t/n}+Kt/n\right)\le\exp[-t].$$

Theorem 8.5 (see [18]). *Let $\mathbf{X}:=(X_1,\ldots,X_n)$. Conditionally on $\mathbf{X}$, for all $t>0$,*

$$\mathbb{P}\left(\mathbf{Z}^\epsilon(\mathcal{F})/C_1\ge\mathbb{E}(\mathbf{Z}^\epsilon(\mathcal{F})|\mathbf{X})+\hat{R}\sqrt{t/n}\,\middle|\,\mathbf{X}\right)\le\exp[-t],$$

where $\hat{R}:=\sup_{f\in\mathcal{F}}\|f\|_n$.

9. Proofs

9.1. Proofs for Section 2

Theorem 2.1 follows from [12]. We present a proof for completeness and to facilitate the extension to products of functions.

Proof of Theorem 2.1. We consider the symmetrized process

$$P_n^\epsilon f^2:=\sum_{i=1}^n\epsilon_i f^2(X_i)/n,$$

with $\epsilon_1,\ldots,\epsilon_n$ a Rademacher sequence independent of $X_1,\ldots,X_n$, and then apply Dudley's theorem, see Theorem 8.3. Note that for two functions f and $\tilde{f}$ in the class $\mathcal{F}$

$$\|f^2-\tilde{f}^2\|_n\le\|f+\tilde{f}\|_n\|f-\tilde{f}\|_{n,\infty}\le 2\hat{R}\|f-\tilde{f}\|_{n,\infty}.$$

It follows that

$$\mathcal{H}(u,\mathcal{F}^2,\|\cdot\|_n)\le\mathcal{H}(u/(2\hat{R}),\mathcal{F},\|\cdot\|_{n,\infty}),\ u>0.$$

Hence

$$\int_\delta^1\sqrt{\mathcal{H}(u\hat{R}K,\mathcal{F}^2,\|\cdot\|_n}du\le\int_\delta^1\sqrt{\mathcal{H}(uK/2,\mathcal{F},\|\cdot\|_{n,\infty})}du.$$

Here we used that $\|f^2\|_n \leq \hat{R}K$. So by Theorem 8.3

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}} P_n^\epsilon f^2\right) \leq C_0 \inf_{\delta>0} \mathbb{E}\left[\hat{R}K \int_\delta^1 \sqrt{\mathcal{H}(uK/2,\mathcal{F},\|\cdot\|_{n,\infty})}du/\sqrt{n} + \delta\hat{R}K\right]$$

$$\leq J_\infty(K,\mathcal{F})\sqrt{\mathbb{E}\hat{R}^2}/\sqrt{n}.$$

But then by Theorem 8.1

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\|f\|_n^2 - \|f\|^2\right|\right) \leq 2J_\infty(K,\mathcal{F})\sqrt{\mathbb{E}\hat{R}^2}/\sqrt{n}. \tag{9.1}$$

This leads to the by-product of the theorem: the inequality

$$\mathbb{E}\hat{R}^2 \leq R^2 + 2J_\infty(K,\mathcal{F})\sqrt{\mathbb{E}\hat{R}^2}/\sqrt{n}$$

gives

$$\sqrt{\mathbb{E}\hat{R}^2} \leq R + 2J_\infty(K,\mathcal{F})/\sqrt{n}.$$

Insert this in (9.1) to find

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\|f\|_n^2 - \|f\|^2\right|\right) \leq 2J_\infty(K,\mathcal{F})R/\sqrt{n} + 4J_\infty^2(K,\mathcal{F})/n.$$

We now apply Theorem 8.4. We have

$$\sup_{f\in\mathcal{F}}\|f^2\| \leq RK, \ \sup_{f\in\mathcal{F}}\|f^2\|_\infty \leq K^2.$$

Hence, inserting the just obtained bound for the expectation, for all $t>0$

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\|f\|_n^2 - \|f\|^2\right|/C_1 \geq \frac{2RJ_\infty(K,\mathcal{F}) + RK\sqrt{t}}{\sqrt{n}} + \frac{4J_\infty^2(K,\mathcal{F}) + K^2 t}{n}\right)$$

$$\leq \exp[-t]. \qquad \square$$

Proof of Theorem 2.2. We start as in the proof of Theorem 2.1 by considering the symmetrized process

$$P_n^\epsilon f^2 := \sum_{i=1}^n \epsilon_i f^2(X_i)/n$$

with $\epsilon_1,\ldots,\epsilon_n$ a Rademacher sequence independent of $X_1,\ldots,X_n$. But when applying Dudley's theorem, see Theorem 8.3, we use a different entropy bound. For two functions f and $\tilde{f}$ in the class $\mathcal{F}$

$$\|f^2 - \tilde{f}^2\|_n \leq \|f+\tilde{f}\|_\infty \|f-\tilde{f}\|_n \leq 2K\|f-\tilde{f}\|_n.$$

It follows that

$$\mathcal{H}(u,\mathcal{F}^2,\|\cdot\|_n) \leq \mathcal{H}(u/(2K),\mathcal{F},\|\cdot\|_n), \ u>0.$$

Hence

$$\int_\delta^1 \sqrt{\mathcal{H}(u\hat{R}K, \mathcal{F}^2, \|\cdot\|_n}du \leq \int_\delta^1 \sqrt{\mathcal{H}(u\hat{R}/2, \mathcal{F}, \|\cdot\|_n)}du.$$

So by Theorem 8.3

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}} P_n^\epsilon f^2\right) \leq K\mathbb{E}\mathcal{J}_0(\hat{R}, \mathcal{F})/\sqrt{n}.$$

Since $v \mapsto \mathcal{J}_0(\sqrt{v}, \mathcal{F})$ is concave

$$\mathbb{E}\mathcal{J}_0(\hat{R}, \mathcal{F}) \leq \mathcal{J}_0(\sqrt{\mathbb{E}\hat{R}^2}, \mathcal{F}).$$

But then by Theorem 8.1

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\|f\|_n^2 - \|f\|^2\right|\right) \leq 2K\mathcal{J}_0(\sqrt{\mathbb{E}\hat{R}^2}, \mathcal{F})/\sqrt{n}. \tag{9.2}$$

This leads to the by-product of the theorem:

$$\mathbb{E}\hat{R}^2 \leq R^2 + 2K\mathcal{J}_0(\sqrt{\mathbb{E}\hat{R}^2}, \mathcal{F})/\sqrt{n} \leq R^2 + \mathbb{E}\hat{R}^2/2 + H(4K/\sqrt{n}).$$

or

$$\mathbb{E}\hat{R}^2 \leq 2R^2 + H(4K/\sqrt{n}) \leq 4R^2.$$

Insert this back to find

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\|f\|_n^2 - \|f\|^2\right|\right) \leq 2K\mathcal{J}_0(2R, \mathcal{F})/\sqrt{n}.$$

Finally apply Theorem 8.4. □

Proof of Theorem 2.3.

Case 1. This follows from

$$\int_0^z \sqrt{H(u, \mathcal{F}, \|\cdot\|_\infty)}du \leq \sqrt{p}\int_0^z \sqrt{H(u/p, \mathcal{F}_0, \|\cdot\|_\infty)}du \asymp \sqrt{p}p^\alpha z^{1-\alpha},$$

and inserting this in Theorem 2.1.

Case 2. Here we use that by conditions (2.7) and (2.4), for $f = \sum_{k=1}^p f_{0,k}$,

$$\|f\|_\infty \leq c_2 \sum_{j=1}^p \|f_{0,k}\|^{1-\alpha} \leq c_2 p^{\frac{1+\alpha}{2}}\left(\sum_{k=1}^p \|f_{0,k}\|^2\right)^{\frac{1-\alpha}{2}} \leq c_2 p^{\frac{1+\alpha}{2}} c_1^{1-\alpha}\|f\|^{1-\alpha}.$$

The result then follows applying the entropy bound of Case 1.

Case 3. For $f(x_1, \ldots, x_p) = \sum_{k=1}^p \sum_{r=1}^N \beta_{r,k}\psi_r(x_k) \in \mathcal{F}$ we have

$$\sum_{k=1}^p \sum_{r=1}^N |\beta_{r,k}| \leq p.$$

The result follows from Theorem 4.1.

Case 4. For $f(x_1,\ldots,x_p)=\sum_{k=1}^p\sum_{r=1}^N\beta_{r,k}\psi_r(x_k)\in\mathcal{F}$, write (with some abuse of notation) $\beta_k:=(\beta_{1,k},\ldots,\beta_{N,k})^T$ and $f_{\beta_k,k}(x_1,\ldots,x_p):=\sum_{r=1}^N\beta_{r,k}(x_k)$. Then by conditions (2.9) and (2.4)

$$\sum_{k=1}^p\|\beta_k\|^2\le c_0\sum_{k=1}^p N\|f_{\beta_k}\|^2\le c_0c_1N\|f\|^2.$$

The result then follows from Theorem 4.3.

Case 5. Let

$$\tilde{\mathcal{F}}:=\Big\{\tilde f(x_1,\ldots,x_p)=\sum_{k=1}^p\sum_{r=1}^N\psi_r(x_k)\Big\}.$$

For all $f\in\mathcal{F}(1)$ with $\|f\|\le 1$ there is a $\tilde f\in\tilde{\mathcal{F}}$ such that

$$\|f-\tilde f\|_\infty\le c_0pN^{-\frac{1}{2\alpha}}.$$

It follows that $\|\tilde f\|\le 1+c_0pN^{-\frac{1}{2\alpha}}\le 2$ for $N\ge(c_0p)^{2\alpha}$. Define

$$\mathbf{Z}_0(\tilde{\mathcal{F}}(1)):=\sup_{\tilde f\in\tilde{\mathcal{F}},\ \|\tilde f\|\le 1}\Big|\|\tilde f\|_n-\|\tilde f\|\Big|.$$

Then

$$\begin{aligned}\mathbf{Z}_0(\tilde{\mathcal{F}}(1))&=\sup_{\tilde f\in\tilde{\mathcal{F}},\ \|\tilde f\|\le 1}\left|\frac{\|\tilde f\|_n-\|\tilde f\|}{\|\tilde f\|}\right|\|\tilde f\|\le\sup_{\tilde f\in\tilde{\mathcal{F}},\ \|\tilde f\|\le 1}\left|\frac{\|\tilde f\|_n-\|\tilde f\|}{\|\tilde f\|}\right|\\&=\sup_{\tilde f\in\tilde{\mathcal{F}},\ \|\tilde f\|=1}\Big|\|\tilde f\|_n-\|\tilde f\|\Big|=\sup_{\tilde f\in\tilde{\mathcal{F}},\ \|\tilde f\|=1}\left|\frac{\|\tilde f\|_n^2-\|\tilde f\|^2}{\|\tilde f\|_n+\|\tilde f\|}\right|\\&\le\sup_{\tilde f\in\tilde{\mathcal{F}},\ \|\tilde f\|=1}\left|\frac{\|\tilde f\|_n^2-\|\tilde f\|^2}{\|\tilde f\|}\right|=\sup_{\tilde f\in\tilde{\mathcal{F}},\ \|\tilde f\|=1}\Big|\|\tilde f\|_n^2-\|\tilde f\|^2\Big|\\&\le\sup_{\tilde f\in\tilde{\mathcal{F}},\ \|\tilde f\|\le 1}\Big|\|\tilde f\|_n^2-\|\tilde f\|^2\Big|:=\mathbf{Z}(\tilde{\mathcal{F}}(1))=\mathcal{O}\Big(N\sqrt{\frac{p\log(pN)}{n}}\Big)\end{aligned}$$

where the last step follows from the same arguments as for Case 4. Define now

$$\mathbf{Z}_0(\mathcal{F}(1)):=\sup_{f\in\mathcal{F},\ \|f\|\le 1}\Big|\|f\|_n-\|f\|\Big|.$$

Clearly

$$\mathbf{Z}_0(\mathcal{F}(1))\le 2\mathbf{Z}_0(\tilde{\mathcal{F}}(1))+2c_0pN^{-\frac{1}{2\alpha}}.$$

So we find

$$\mathbf{Z}_0(\mathcal{F}(1))=\mathcal{O}_{\mathbb{P}}(1)\underbrace{\Big(N\sqrt{\frac{p\log(pN)}{n}}\Big)}_{I}+\underbrace{2c_0pN^{-\frac{1}{2\alpha}}}_{II}.$$

Since $p \le n$ choosing $N \asymp (np/\log(n))^{\frac{\alpha}{(1+2\alpha)}}$ gives

$$I \asymp II \asymp p^{\frac{1+4\alpha}{2(1+2\alpha)}} \Big(\log n/n\Big)^{\frac{1}{2(1+2\alpha)}},$$

so that $\mathbf{Z}_0(\mathcal{F}(1)) = \mathcal{O}_{\mathbb{P}}(p^{\frac{1+4\alpha}{2(1+2\alpha)}}(\log n/n)^{\frac{1}{2(1+2\alpha)}})$. But then also

$$\mathbf{Z}(\mathcal{F}(1)) = \sup_{f\in\mathcal{F},\ \|f\|\le 1} \Big| \|f\|_n^2 - \|f\|^2 \Big|$$

$$\le 2 \sup_{f\in\mathcal{F},\ \|f\|\le 1} \Big| \|f\|_n - \|f\| \Big| + \sup_{f\in\mathcal{F},\ \|f\|\le 1} \Big| \|f\|_n - \|f\| \Big|^2$$

$$= 2\mathbf{Z}_0(\mathcal{F}(1)) + \mathbf{Z}_0^2(\mathcal{F}(1)) = \mathcal{O}_{\mathbb{P}}(p^{\frac{1+4\alpha}{2(1+2\alpha)}}(\log n/n)^{\frac{1}{2(1+2\alpha)}}). \qquad \square$$

Proof of Lemma 2.1. Let $N \in \mathbb{N}$ be arbitrary and let for $f_0 \in \mathcal{F}_0$ and certain β_r and ψ_r, $r = 1, \dots N$:

$$\|f_0 - \sum_{r=1}^{N} \beta_r \psi_r\|_\infty \le c_0 N^{-\frac{1}{2\alpha}}.$$

Let $\beta := (\beta_1, \dots, \beta_N)^T$, $f_\beta := \sum_{r=1}^N \beta_r \psi_r$ and $f_{\beta,k}(x_1, \dots, x_p) := f_\beta(x_k)$ and let $f_{0,k}(x_1, \dots, x_p) := f_0(x_k)$. Then for $N \ge \|f_{0,k}\|^{-2\alpha}$

$$\|f_{\beta,k}\| \le \|f_{0,k}\| + c_0 N^{-\frac{1}{2\alpha}} \le (1 + c_0)\|f_{0,k}\|.$$

Define $K_\psi := \sup_r \|\psi_r\|_\infty \vee 1$. We find for $N \ge 1$, inserting (2.10),

$$\|f_0\|_\infty \le \|\sum_{r=1}^{N} \beta_r \psi_r\|_\infty + c_0 N^{-\frac{1}{2\alpha}}$$

$$\le (1 + c_3)\|\beta\|_\infty K_\psi + c_0 N^{-\frac{1}{2\alpha}} \le (1 + c_3)\|\beta\|_2 K_\psi + c_0\|f_{0,k}\|$$

$$\le c_0(1 + c_3)K_\psi \sqrt{N}\|f_{\beta,k}\| + c_0\|f_{0,k}\| \le c_0(2 + c_3)K_\psi \sqrt{N}\|f_{0,k}\|$$

where in the second last inequality we used (2.9). Take N as the smallest integer greater than or equal to $\|f_{0,k}\|^{-2\alpha}$. Then $N \le \|f_{0,k}\|^{-2\alpha} + 1$ so that $\sqrt{N} \le \|f_{0,k}\|^{-\alpha} + 1$ and hence

$$\|f_0\|_\infty \le c_0(2 + c_3)K_\psi(\|f_{0,k}\|^{1-\alpha} + \|f_{0,k}\|) \le 2c_0(2 + c_3)^2 K_\psi \|f_{0,k}\|^{1-\alpha},$$

since by (2.4) $\|f_{0,k}\| \le 1$. Hence (2.7) holds with $c_2 = 2c_0(2 + c_3)^2 K_\psi$. $\square$

9.2. Proofs for Section 3

Proof of Theorem 3.1. Let

$$\hat{R}_1 := \sup_{f \in \mathcal{F}} \|f\|_n, \ \hat{R}_2 := \sup_{g \in \mathcal{G}} \|g\|_n.$$

For functions $f, \tilde{f}$ in the class $\mathcal{F}$ and $g, \tilde{g}$ in the class $\mathcal{G}$ we have

$$\|fg - \tilde{f}\tilde{g}\|_n \leq \|fg - \tilde{f}g\|_n + \|\tilde{f}g - \tilde{f}\tilde{g}\|_n \leq \hat{R}_2 \|f - \tilde{f}\|_{n,\infty} + \hat{R}_1 \|g - \tilde{g}\|_{n,\infty}.$$

It follows that

$$\mathcal{H}(u, \mathcal{F} \times \mathcal{G}, \|\cdot\|_n)$$
$$\leq \mathcal{H}(u/(2\hat{R}_2), \mathcal{F}, \|\cdot\|_{n,\infty}) + \mathcal{H}(u/(2\hat{R}_1), \mathcal{G}, \|\cdot\|_{n,\infty}), \ u > 0.$$

We moreover have

$$\|fg\|_n \leq (\hat{R}_2 K_1) \wedge (\hat{R}_1 K_2) \leq \hat{R}_1 K_2.$$

Define the set

$$\mathcal{R} := \{\hat{R}_1 \leq 2R_1, \ \hat{R}_2 \leq 2R_2\}.$$

By Theorem 2.1, and since $J_\infty \leq \mathcal{J}_\infty$, for values of t and n satisfying (3.1) and (3.2) it holds that

$$\mathbb{P}(\mathcal{R}) \geq 1 - 2\exp[-t].$$

On $\mathcal{R}$

$$\int_{\delta \hat{R}_1 K_2}^{\hat{R}_1 K_2} \sqrt{\mathcal{H}(u, \mathcal{F} \times \mathcal{G}, \|\cdot\|_n)} du$$

$$\leq 2R_1 \int_{\delta K_2}^{K_2} \sqrt{\mathcal{H}(u/2, \mathcal{G}, \|\cdot\|_n)} du + 2R_2 \int_{\delta R_1 K_2/(4R_2)}^{K_2 R_1/R_2} \sqrt{\mathcal{H}(u/2, \mathcal{F}, \|\cdot\|_n)} du.$$

Consider now the symmetrized process

$$P_n^\epsilon fg := \sum_{i=1}^n \epsilon_i f(X_i) g(X_i)/n,$$

with $\epsilon_1, \ldots, \epsilon_n$ a Rademacher sequence independent of $X_1, \ldots, X_n$. By Theorem 8.3 we have now found that conditionally on $\mathbf{X} := \{X_1, \ldots, X_n\}$,

$$\mathbb{E}\Big(\sup_{f \in \mathcal{F}, \ g \in \mathcal{G}} |P_n^\epsilon fg| \Big| \mathbf{X} \Big) \{\mathbf{X} \in \mathcal{R}\}$$

$$\leq 2R_1 \mathcal{J}_\infty(K_2, \mathcal{G})/\sqrt{n} + 2R_2 \mathcal{J}_\infty(R_1 K_2/R_2, \mathcal{F})/\sqrt{n} := \mathbf{E}$$

From Theorem 8.4, we get that conditionally on $\mathbf{X}$

$$\mathbb{P}\Big(\sup_{f \in \mathcal{F}, \ g \in \mathcal{G}} |P_n^\epsilon fg|/C_1 \geq \mathbf{E} + 2R_1 K_2 \sqrt{\frac{t}{n}} + K_1 K_2 \frac{t}{n} \Big| \mathbf{X} \Big) \{\mathbf{X} \in \mathcal{R}\} \leq \exp[-t].$$

But then, since $\mathbb{P}(\mathcal{R}^c) \leq 2\exp[-t]$,

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F},\ g\in\mathcal{G}} |P_n^\varepsilon fg|/C_1 \geq \mathbf{E} + 2R_1K_2\sqrt{\frac{t}{n}} + K_1K_2\frac{t}{n}\Big) \leq 3\exp[-t].$$

Now apply Theorem 8.2. □

Proof of Theorem 3.2. Let $\{\varepsilon_i\}_{i=1}^n$ be a Rademacher sequence independent of $\{X_i, Y_i\}_{i=1}^n$. Conditionally on $(\mathbf{X}, \mathbf{Y}) := (\{X_1, \ldots, X_n\}, \{Y_1, \ldots, Y_n\})$, by Theorem 8.3, for $P_n^\varepsilon \mathbf{Y} f := \sum_{i=1}^n \varepsilon_i Y_i f(X_i)/n$, $f \in \mathcal{F}$:

$$\mathbb{E}\Big(\sup_{f\in\mathcal{F}} |P_n^\varepsilon \mathbf{Y} f| \Big| \mathbf{X}, \mathbf{Y}\Big) \leq C_0 \mathcal{J}_\infty(\|\mathbf{Y}\|_n K, \{Yf : \ f \in \mathcal{F}\}).$$

So on the set

$$\mathcal{Y} := \{\|\mathbf{Y}\|_n^2 \leq 4K_0^2\}$$

we get

$$\mathbb{E}\Big(\sup_{f\in\mathcal{F}} |P_n^\varepsilon \mathbf{Y} f| \Big| \mathbf{X}, \mathbf{Y}\Big)\{\mathbf{Y} \in \mathcal{Y}\} \leq 2\mathcal{J}_\infty(K_0K, \mathcal{F})/\sqrt{n}.$$

Now apply Theorem 8.5 to obtain that for all $t > 0$

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}} |P_n^\varepsilon \mathbf{Y} f|/C_1 \geq \frac{\mathcal{J}_\infty(KK_0, \mathcal{F})}{\sqrt{n}} + 2KK_0R\sqrt{\frac{t}{n}} \Big| \mathbf{X}, \mathbf{Y}\Big)\{\mathbf{Y} \in \mathcal{Y}\} \leq \exp[-t].$$

We now integrate out and use that $\mathbb{P}(\mathcal{Y}^c) \leq \exp[-t]$. Then we de-symmetrize using Theorem 8.2. □

9.3. Proofs for Section 4

Proof of Theorem 4.1. If $\|\beta\|_1 \leq M$ we know that

$$\|f_\beta\|_\infty \leq \|\beta\|_1 \max_{1\leq i\leq n} \max_{1\leq j\leq p} |X_{i,j}| \leq MK_X.$$

Fixing δ at $\delta = 1/\sqrt{n}$ in (2.1), we find by Lemma 4.1

$$J(\mathcal{F}) \leq C_0 MK_X \int_{1/\sqrt{n}}^1 \sqrt{\mathcal{H}(u/2, \{f_\beta : \ \|\beta\|_1 \leq M\}, \|\cdot\|_{n,\infty})}du + C_0MK$$

$$= 2C_0\sqrt{\log(2p)\log(2n)}MK_X \log n + C_0MK_X.$$

The result now follows from Theorem 2.1. □

Proof of Theorem 4.2. Since $\|f_{\beta_S}\| \leq 1$ implies

$$\|\beta_S\|_1 \leq \sqrt{s}\|\beta\|_2 \leq \sqrt{s}/\Lambda_{\min}$$

this follows directly from Theorem 4.1. □

Proof of Lemma 4.2. We let $\psi_j(X_i) := X_{i,j}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$. It holds that

$$\sup_{\|f_\beta\| \le 1} |(P_n - P)\mathbf{Y} f_\beta|^2 \le \sup_{\|f_\beta\| \le 1} \sum_{j=1}^p |(P_n - P)\mathbf{Y}\psi_j|^2 \|\beta\|_2^2$$

$$\le \sum_{j=1}^p |(P_n - P)\mathbf{Y}\psi_j|^2 / \Lambda_{\min}^2.$$

But by the triangle inequality

$$\left\| \sum_{j=1}^p |(P_n - P)\mathbf{Y}\psi_j|^2 \right\|_{\Psi_1} \le \sum_{j=1}^p \left\| |(P_n - P)\mathbf{Y}\psi_j|^2 \right\|_{\Psi_1}.$$

Moreover for all j,

$$\left\| |(P_n - P)\mathbf{Y}\psi_j|^2 \right\|_{\Psi_1} = \left\| (P_n - P)\mathbf{Y}\psi_j \right\|_{\Psi_2}^2 \le c_2^2 K_0^2 K_X^2 / n.$$

Hence

$$\left\| \sqrt{\sum_{j=1}^p |(P_n - P)\mathbf{Y}\psi_j|^2} \right\|_{\Psi_2} = \sqrt{\left\| \sum_{j=1}^p |(P_n - P)\mathbf{Y}\psi_j|^2 \right\|_{\Psi_1}} \le c_2 K_0 K_X \sqrt{\frac{p}{n}}.$$

By Chebyshev's inequality, for all $t > 0$,

$$\mathbb{P}\left(\sqrt{\sum_{j=1}^p |(P_n - P)\mathbf{Y}\psi_j|^2} \ge \frac{c_2 K_0 K_X}{\Lambda_{\min}} \sqrt{\frac{pt}{n}} \right) \le 2\exp[-t]. \qquad \square$$

9.4. Proofs for Section 5

Proof of Lemma 5.1. The inequality

$$\|\mathbf{Y} - \hat{f}\|_n \le \|\mathbf{Y} - f^*\|_n,$$

can be rewritten to the Basic Inequality

$$\|\hat{f} - f^*\|_n^2 \le 2P_n(\mathbf{Y} - f^*)(\hat{f} - f^*).$$

On $\mathcal{T}$ we therefore have

$$\|\hat{f} - f^*\|^2 \le \|\hat{f} - f^*\|^2 - \|\hat{f} - f^*\|_n^2 + 2P_n(\mathbf{Y} - f^*)(\hat{f} - f^*)$$

$$\le \delta_n \|\hat{f} - f^*\|^2 + \delta_n \|\hat{f} - f^*\|,$$

where we used that $P(\mathbf{Y} - f^*)(\hat{f} - f^*) = 0$. Hence

$$\|\hat{f} - f^*\| \le \delta_n / (1 - \delta_n) \le 2\delta_n. \qquad \square$$

Proof of Theorem 5.1. This follows from Lemma 5.1 combined with Summary 4.1. We use here that $\|f^*\| \le \|\mathbf{Y}\| \le K_0$. The first result then follows immediately from Lemma 5.1 and Summary 4.1. For the second result, write

$$\|\mathbf{Y} - \hat{f}\|_n^2 - \|\mathbf{Y} - f^*\|_n^2 = \|\hat{f} - f^*\|_n^2 - 2P_n(\mathbf{Y} - f^*)(\hat{f} - f^*).$$

But

$$\|\hat{f} - f^*\|_n^2 = \|\hat{f} - f^*\|^2 + \left(\frac{\|\hat{f} - f^*\|_n^2}{\|\hat{f} - f^*\|^2} - 1\right)\|\hat{f} - f^*\|^2$$
$$= \mathcal{O}_{\mathbb{P}}(\delta_n^2 + \delta_n^3)$$

and

$$P_n(\mathbf{Y} - f^*)(\hat{f} - f^*) = (P_n - P)(\mathbf{Y} - f^*)(\hat{f} - f^*)$$
$$= \|\hat{f} - f^*\|\left((P_n - P)(\mathbf{Y} - f^*)(\hat{f} - f^*)/\|\hat{f} - f^*\|\right) = O_{\mathbb{P}}(\delta_n^2).$$

Hence

$$\|\mathbf{Y} - \hat{f}\|_n^2 - \|\mathbf{Y} - f^*\|^2 = \|\mathbf{Y} - \hat{f}\|_n^2 - \|\mathbf{Y} - f^*\|_n^2 + \|\mathbf{Y} - f^*\|_n^2 + \|\mathbf{Y} - f^*\|_n^2$$
$$= O_{\mathbb{P}}(\delta_n^2) + \|\mathbf{Y} - f^*\|_n^2 - \|\mathbf{Y} - f^*\|^2.$$

We find

$$\|\mathbf{Y} - f^*\|_n^2 - \|\mathbf{Y} - f^*\|^2 = \|\mathbf{Y}\|_n^2 - \|\mathbf{Y}\|^2 - 2(P_n - P)(\mathbf{Y}f^*) + \|f^*\|_n^2 - \|f^*\|^2$$
$$= \mathcal{O}_{\mathbb{P}}(1/\sqrt{n}) + \mathcal{O}_{\mathbb{P}}(\delta_n). \qquad \square$$

Proof of Theorem 5.2. By Summary 4.1 all probability statements are uniformly in S, so that the set $\mathcal{T}$ given in Lemma 5.1 has with $\delta_n = O(\sqrt{p \log p/n})$ the required large probability. $\square$

9.5. Proof for Section 6

Proof of Theorem 6.1. From Theorem 2.3

$$\sup_{f \in \mathcal{F},\ \|f\| \le 1} \left| \|f\|_n^2 - \|f\|^2 \right| = o_{\mathbb{P}}(1).$$

We know moreover from Lemma 3.2 that also for

$$\max_j \sup_{f \in \mathcal{F},\ \|f\| \le 1} (P_n - P)\mathbf{X}_j f = o_{\mathbb{P}}(1).$$

Also

$$\max_j \left| \|\mathbf{X}_j\|_n^2 - \|X_j\|^2 \right| = o_{\mathbb{P}}(1).$$

Hence,

$$\max_j \sup_{f \in \mathcal{F},\ \|f\| \le 1} \left| \|\mathbf{X}_j - f\|_n^2 - \|\mathbf{X}_j - f\|^2 \right| = o_{\mathbb{P}}(1). \tag{9.3}$$

We now note that we only need uniform convergence over $f \in \mathcal{F}$ with $\|f\| \leq 1$. To see this, let for any π and j

$$\tilde{f}_j(\pi) := s\hat{f}_j(\pi) + (1-s)f_j^*(\pi)$$

where $s := c/(c + \|\hat{f}_j(\pi) - f_j^*(\pi)\|)$ and c a constant (depending on j and π) to be chosen (see below). Then $\|\tilde{f}_j(\pi) - f_j^*(\pi)\| \leq 1$. Moreover

$$\|\mathbf{X}_j - \tilde{f}_j(\pi)\|_n^2 \leq s\|\mathbf{X}_j - \hat{f}_j(\pi)\|_n^2 + (1-s)\|\mathbf{X}_j - f_j^*(\pi)\|_n^2 \leq \|\mathbf{X}_j - f_j^*(\pi)\|_n^2.$$

So

$$\begin{aligned}\|\mathbf{X}_j - \tilde{f}_j(\pi)\| &= \|\mathbf{X}_j - \tilde{f}_j(\pi)\| - \|\mathbf{X}_j - \tilde{f}_j(\pi)\|_n + \|\mathbf{X}_j - \tilde{f}_j(\pi)\|_n \\ &= \|\mathbf{X}_j - \tilde{f}_j(\pi)\|_n + o_{\mathbb{P}}(1) \leq \|\mathbf{X}_j - f_j^*(\pi)\|_n + o_{\mathbb{P}}(1) \\ &= \|\mathbf{X}_j - f_j^*(\pi)\| + o_{\mathbb{P}}(1) \leq \|\mathbf{X}_j - \tilde{f}_j(\pi)\| + o_{\mathbb{P}}(1)\end{aligned}$$

where in the last step we used the convexity of $\mathcal{F}_0$. Thus $\|\mathbf{X}_j - \tilde{f}_j(\pi)\| = o_{\mathbb{P}}(1)$. This impies that $\|\tilde{f}_j(\pi) - f_j^*(\pi)\| \leq \|\mathbf{X}_j - f_j^*(\pi)\| + o_{\mathbb{P}}(1)$. Choosing c appropriately, for example $c = 4\|\mathbf{X}_j - f_j^*(\pi)\|$, we now find that $\|\hat{f}_j(\pi) - f_j^*(\pi)\| = \mathcal{O}_{\mathbb{P}}(1)$. By applying the same arguments as above with $\tilde{f}_j(\pi)$ replaced by $\hat{f}_j(\pi)$ shows that $\|\mathbf{X}_j - \hat{f}_j(\pi)\| = o_{\mathbb{P}}(1)$. This result is uniformly in $\pi \in \Pi$ by the same arguments as used for Theorem 5.2. Application of the union bound and deviation bounds for each j, we see that the result is also uniformly in j. □

References

[1] Adamczak, R., Litvak, A.E., Pajor, A., and Tomczak-Jaegermann, N., Sharp bounds on the rate of convergence of the empirical covariance matrix. *Comptes Rendus Mathematique*, 349(3):195–200, 2011. MR2769907

[2] Agmon, S. and Jones, F., *Lectures on elliptic boundary value problems Elliptic boundary value problems Van Nostrand mathematical studies*. Van Nostrand, 1965. MR0178246

[3] Ahlswede, R. and Winter, A., Strong converse for identification via quantum channels. *Information Theory, IEEE Transactions on*, 48(3):569–579, 2002. MR1889969

[4] Bartlett, P., Bousquet, O., and Mendelson, S., Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. MR2166554

[5] Bartlett, P.L., Mendelson, S., and Neeman, J., ℓ_1-regularized linear regression: persistence and oracle inequalities. *Probability Theory and Related Fields*, 154(1–2):193–224, 2012. MR2981422

[6] Bickel, P., Ritov, Y., and Tsybakov, A., Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009. MR2533469

[7] Birman, M.Š. and Solomjak, M.Z., Piecewise-polynomial approximations of functions of the classes W_p^α. *Mathematics of the USSR-Sbornik*, 2:295–317, 1967. MR0217487

[8] Bühlmann, P., Peters, J., and Ernest, J., CAM: Causal Additive Models, high-dimensional order search and penalized regression, 2013. ArXiv: 1310.1533. MR3072790

[9] Chen, S.S., Donoho, D.L., and Saunders, M.A., Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998. MR1639094

[10] Dudley, R.M., The sizes of compact subsets of hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967. MR0220340

[11] Giné, E. and Koltchinskii, V., Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006. MR2243881

[12] Guédon, O., Mendelson, S., Pajor, A., and Tomczak-Jaegermann, N., Subspaces and orthogonal decompositions generated by bounded orthogonal systems. *Positivity*, 11(2):269–283, 2007. MR2321621

[13] Koltchinskii, V., A remark on low rank matrix recovery and noncommutative Bernstein type inequalities. In *IMS Collections From Probability to Statistics and Back: High-Dimensional Models and Processes*, volume 9, pages 213–226. Institute of Mathematical Statistics, Beachwood, Ohio, 2013. Banerjee, M., Bunea, F., Huang, J., Koltchinskii, V., and Maathuis, M.H., eds.

[14] Lecué, G. and Mendelson, S., Performance of empirical risk minimization in linear aggregation, 2014. ArXiv: 1402.5763.

[15] Ledoux, M. and Talagrand, M., *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Verlag, New York, 1991. MR1102015

[16] Loh, P.-L. and Wainwright, M.J., High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 40:1637–1664, 2012. MR3015038

[17] Lugosi, G. and Nobel, A., Adaptive model selection using empirical complexities. *Annals of Statistics*, 27:1830–1864, 1999. MR1765619

[18] Massart, P., About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 28:863–884, 2000. MR1782276

[19] Meier, L., Van de Geer, S., and Bühlmann, P., High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009. MR2572443

[20] Müller, P. and van de Geer, S.A., The partial linear model in high dimensions, 2013. Submitted. ArXiv: 1307.1067.

[21] Nickl, R. and Pötscher, B., Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov-and Sobolev-type. *Journal of Theoretical Probability*, 20(2):177–199, 2007. MR2324525

[22] Nickl, R. and van de Geer, S.A., Confidence sets in sparse regression. *The Annals of Statistics*, 41:2852–2876, 2013. MR3161450

[23] Pollard, D., *Convergence of Stochastic Processes*. Springer, 1984. MR0762984

[24] Raskutti, G., Wainwright, M.J., and Yu, B., Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010. MR2719855

[25] Rudelson, M. and Vershynin, R., On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008. MR2417886

[26] Rudelson, M. and Zhou, S., Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59:3434–3447, 2013. MR3061256

[27] Talagrand, M., Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'IHES*, 81:73–205, 1995. MR1361756

[28] van de Geer, S., *Empirical Processes in M-Estimation*. Cambridge University Press, 2000. MR1739079

[29] van de Geer, S., The deterministic Lasso. *The JSM Proceedings*, 2007.

[30] van de Geer, S. and Bühlmann, P., On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, pages 1360–1392, 2009. MR2576316

[31] van de Geer, S.A., *Empirical Processes in M-Estimation*. Cambridge, 2000.

[32] van de Geer, S.A. and Bühlmann, P., ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41:536–567, 2013. MR3099113

[33] van de Geer, S.A. and Muro, A., The additive model with different smoothness for the components, 2014. In progress.

[34] van de Geer, S.A., Bühlmann, P., Ritov, Y., and Dezeure, R., On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 2014. To appear.

[35] van der Vaart, A.W. and Wellner, J.A., *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3. MR1385671