Journal of Machine Learning Research 13 (2012) 2465-2501 Submitted 3/11; Revised 2/12; Published 8/12

On the Convergence Rate of ℓ_p-Norm Multiple Kernel Learning*

Marius Kloft† KLOFT@TU-BERLIN.DE
Machine Learning Laboratory
Technische Universität Berlin
Franklinstr. 28/29
10587 Berlin, Germany

Gilles Blanchard GILLES.BLANCHARD@MATH.UNI-POTSDAM.DE
Department of Mathematics
University of Potsdam
Am Neuen Palais 10
14469 Potsdam, Germany

Editor: Sören Sonnenburg, Francis Bach, and Cheng Soog Ong

Abstract

We derive an upper bound on the local Rademacher complexity of ℓ_p-norm multiple kernel learning, which yields a tighter excess risk bound than global approaches. Previous local approaches analyzed the case $p = 1$ only while our analysis covers all cases $1 \leq p \leq \infty$, assuming the different feature mappings corresponding to the different kernels to be uncorrelated. We also show a lower bound that shows that the bound is tight, and derive consequences regarding excess loss, namely fast convergence rates of the order $O(n^{-\frac{\alpha}{1+\alpha}})$, where α is the minimum eigenvalue decay rate of the individual kernels.

Keywords: multiple kernel learning, learning kernels, generalization bounds, local Rademacher complexity

1. Introduction

Propelled by the increasing "industrialization" of modern application domains such as bioinformatics or computer vision leading to the accumulation of vast amounts of data, the past decade experienced a rapid professionalization of machine learning methods. Sophisticated machine learning solutions such as the support vector machine can nowadays almost completely be applied out-of-the-box (Bouckaert et al., 2010). Nevertheless, a displeasing stumbling block towards the complete automatization of machine learning remains that of finding the best abstraction or *kernel* (Schölkopf et al., 1998; Müller et al., 2001) for a problem at hand.

In the current state of research, there is little hope that in the near future a machine will be able to automatically engineer the *perfect* kernel for a particular problem (Searle, 1980). However, by restricting to a less general problem, namely to a finite set of base kernels the algorithm can pick

∗. This is a longer version of a short conference paper entitled *The Local Rademacher Complexity of ℓ_p-Norm MKL*, which is appearing in Advances in Neural Information Processing Systems 24 edited by J. Shawe-Taylor and R.S. Zemel and P. Bartlett and F. Pereira and K.Q. Weinberger (2011).

†. Parts of the work were done while MK was at Learning Theory Group, Computer Science Division and Department of Statistics, University of California, Berkeley, CA 94720-1758, USA.

from, one might hope to achieve automatic kernel selection: clearly, cross-validation based model selection (Stone, 1974) can be applied if the number of base kernels is decent. Still, the performance of such an algorithm is limited by the performance of the best kernel in the set.

In the seminal work of Lanckriet et al. (2004) it was shown that it is computationally feasible to simultaneously learn a support vector machine *and* a linear combination of kernels at the same time, if we require the so-formed kernel combinations to be positive definite and trace-norm normalized. Though feasible for small sample sizes, the computational burden of this so-called *multiple kernel learning* (MKL) approach is still high. By further restricting the multi-kernel class to only contain convex combinations of kernels, the efficiency can be considerably improved, so that ten thousands of training points and thousands of kernels can be processed (Sonnenburg et al., 2006).

However, these computational advances come at a price. Empirical evidence has accumulated showing that sparse-MKL optimized kernel combinations rarely help in practice and frequently are to be outperformed by a regular SVM using an unweighted-sum kernel $K = \sum_m K_m$ (Cortes et al., 2008; Gehler and Nowozin, 2009), leading for instance to the provocative question "Can learning kernels help performance?" (Cortes, 2009).

A first step towards a model of learning the kernel that is useful in practice was achieved in Kloft et al. (2008), Cortes et al. (2009), Kloft et al. (2009) and Kloft et al. (2011), where an ℓ_q-norm, $q \geq 1$, rather than an ℓ_1 penalty was imposed on the kernel combination coefficients. The ℓ_q-norm MKL is an empirical minimization algorithm that operates on the multi-kernel class consisting of functions $f : x \mapsto \langle \boldsymbol{w}, \phi_k(x) \rangle$ with $\|\boldsymbol{w}\|_k \leq D$, where ϕ_k is the kernel mapping into the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ with kernel k and norm $\|.\|_k$, while the kernel k itself ranges over the set of possible kernels $\left\{ k = \sum_{m=1}^M \theta_m k_m \;\middle|\; \|\boldsymbol{\theta}\|_q \leq 1,\ \boldsymbol{\theta} \geq 0 \right\}$.

In Figure 1, we reproduce exemplary results taken from Kloft et al. (2009, 2011) (see also references therein for further evidence pointing in the same direction). We first observe that, as expected, ℓ_q-norm MKL enforces strong sparsity in the coefficients θ_m when $q = 1$, and no sparsity at all for $q = \infty$, which corresponds to the SVM with an unweighted-sum kernel, while intermediate values of q enforce different degrees of soft sparsity (understood as the steepness of the decrease of the ordered coefficients θ_m). Crucially, the performance (as measured by the AUC criterion) is not monotonic as a function of q; $q = 1$ (sparse MKL) yields significantly worse performance than $q = \infty$ (regular SVM with sum kernel), but optimal performance is attained for some intermediate value of q. This is an empirical strong motivation to theoretically study the performance of ℓ_q-MKL beyond the limiting cases $q = 1$ or $q = \infty$.

A conceptual milestone going back to the work of Bach et al. (2004) and Micchelli and Pontil (2005) is that the above multi-kernel class can equivalently be represented as a block-norm regularized linear class in the product Hilbert space $\mathcal{H} := \mathcal{H}_1 \times \cdots \times \mathcal{H}_M$, where $\mathcal{H}_m$ denotes the RKHS associated to kernel k_m, $1 \leq m \leq M$. More precisely, denoting by ϕ_m the kernel feature mapping associated to kernel k_m over input space $\mathcal{X}$, and $\phi : x \in \mathcal{X} \mapsto (\phi_1(x), \ldots, \phi_M(x)) \in \mathcal{H}$, the class of functions defined above coincides with

$$H_{p,D,M} = \left\{ f_{\boldsymbol{w}} : x \mapsto \langle \boldsymbol{w}, \phi(x) \rangle \;\middle|\; \boldsymbol{w} = (\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(M)}), \|\boldsymbol{w}\|_{2,p} \leq D \right\}, \tag{1}$$

where there is a one-to-one mapping of $q \in [1, \infty]$ to $p \in [1,2]$ given by $p = \frac{2q}{q+1}$ (see Appendix A for a derivation). The $\ell_{2,p}$-norm is defined here as $\left\|\boldsymbol{w}\right\|_{2,p} := \left\| \left(\|\boldsymbol{w}^{(1)}\|_{k_1}, \ldots, \|\boldsymbol{w}^{(M)}\|_{k_M} \right) \right\|_p = \left(\sum_{m=1}^M \left\|\boldsymbol{w}^{(m)}\right\|_{k_m}^p \right)^{1/p}$; for simplicity, we will frequently write $\left\|\boldsymbol{w}^{(m)}\right\|_2 = \left\|\boldsymbol{w}^{(m)}\right\|_{k_m}$.

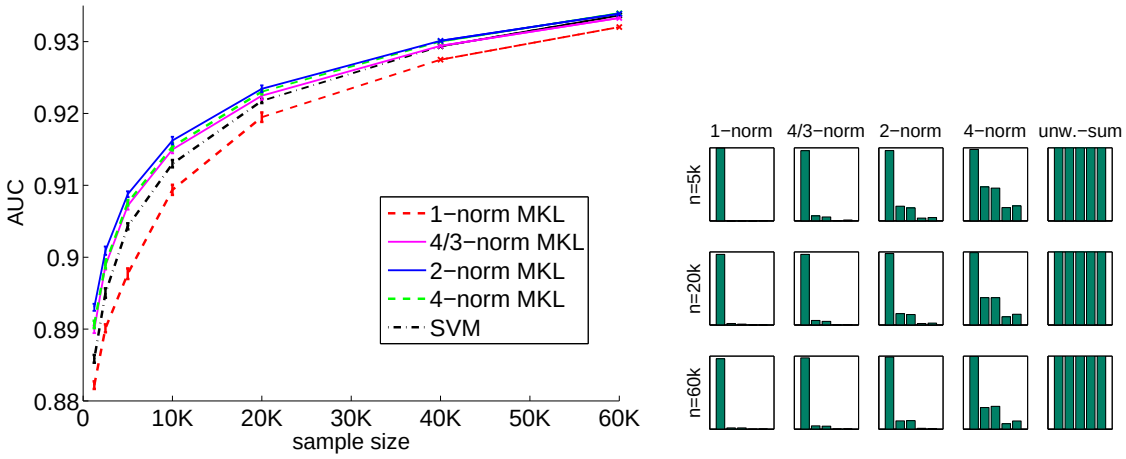


Figure 1: Splice site detection experiment in Kloft et al. (2009, 2011). LEFT: The Area under ROC curve as a function of the training set size is shown. The regular SVM is equivalent to $q = \infty$ (or $p = 2$). RIGHT: The optimal kernel weights θ_m as output by ℓ_q-norm MKL are shown.

Clearly, the complexity of the class (1) will be greater than one that is based on a single kernel only. However, it is unclear whether the increase is decent or considerably high and—since there is a free parameter p—how this relates to the choice of p. To this end the main aim of this paper is to analyze the sample complexity of the above hypothesis class (1). An analysis of this model, based on global Rademacher complexities, was developed by Cortes et al. (2010). In the present work, we base our main analysis on the theory of *local* Rademacher complexities, which allows to derive improved and more precise rates of convergence.

1.1 Outline of the Contributions

This paper makes the following contributions:

- Upper bounds on the local Rademacher complexity of ℓ_p-norm MKL are shown, from which we derive an excess risk bound that achieves a fast convergence rate of the order $O(M^{1+\frac{2}{1+\alpha}\left(\frac{1}{p^*}-1\right)} n^{-\frac{\alpha}{1+\alpha}})$, where α is the minimum eigenvalue decay rate of the individual kernels[1] (previous bounds for ℓ_p-norm MKL only achieved $O(M^{\frac{1}{p^*}} n^{-\frac{1}{2}})$.
- A lower bound is shown that besides absolute constants matches the upper bounds, showing that our results are tight.
- The generalization performance of ℓ_p-norm MKL as guaranteed by the excess risk bound is studied for varying values of p, shedding light on the appropriateness of a small/large p in various learning scenarios.

1. That is, it $\exists d > 0$ and $\alpha > 1$ such that for all $m = 1, \ldots, M$ it holds $\lambda_j^{(m)} \leq d j^{-\alpha}$, where $\lambda_j^{(m)}$ is the jth eigenvalue of the mth kernel (sorted in descending order).

Furthermore, we also present a simple proof of a global Rademacher bound similar to the one shown in Cortes et al. (2010). A comparison of the rates obtained with local and global Rademacher analysis, respectively, can be found in Section 6.1.

1.2 Notation

For notational simplicity we will omit feature maps and directly view $\phi(x)$ and $\phi_m(x)$ as random variables $\boldsymbol{x}$ and $\boldsymbol{x}^{(m)}$ taking values in the Hilbert space $\mathcal{H}$ and $\mathcal{H}_m$, respectively, where $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)})$. Correspondingly, the hypothesis class we are interested in reads $H_{p,D,M} = \left\{ f_{\boldsymbol{w}} : \boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x} \rangle \;\middle|\; \|\boldsymbol{w}\|_{2,p} \leq D \right\}$. If D or M are clear from the context, we sometimes synonymously denote $H_p = H_{p,D} = H_{p,D,M}$. We will frequently use the notation $(\boldsymbol{u}^{(m)})_{m=1}^M$ for the element $\boldsymbol{u} = (\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)}) \in \mathcal{H} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_M$.

We denote the kernel matrices corresponding to k and k_m by K and K_m, respectively. Note that we are considering normalized kernel Gram matrices, that is, the ijth entry of K is $\frac{1}{n} k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. We will also work with covariance operators in Hilbert spaces. In a finite dimensional vector space, the (uncentered) covariance operator can be defined in usual vector/matrix notation as $\mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top$. Since we are working with potentially infinite-dimensional vector spaces, we will use instead of $\boldsymbol{x}\boldsymbol{x}^\top$ the tensor notation $\boldsymbol{x} \otimes \boldsymbol{x} \in \mathrm{HS}(\mathcal{H})$, which is a Hilbert-Schmidt operator $\mathcal{H} \mapsto \mathcal{H}$ defined as $(\boldsymbol{x} \otimes \boldsymbol{x})\boldsymbol{u} = \langle \boldsymbol{x}, \boldsymbol{u} \rangle \boldsymbol{x}$. The space $\mathrm{HS}(\mathcal{H})$ of Hilbert-Schmidt operators on $\mathcal{H}$ is itself a Hilbert space, and the expectation $\mathbb{E}\boldsymbol{x} \otimes \boldsymbol{x}$ is well-defined and belongs to $\mathrm{HS}(\mathcal{H})$ as soon as $\mathbb{E}\|\boldsymbol{x}\|^2$ is finite, which will always be assumed (as a matter of fact, we will often assume that $\|\boldsymbol{x}\|$ is bounded a.s.). We denote by $J = \mathbb{E}\boldsymbol{x} \otimes \boldsymbol{x}$, $J_m = \mathbb{E}\boldsymbol{x}^{(m)} \otimes \boldsymbol{x}^{(m)}$ the uncentered covariance operators corresponding to variables $\boldsymbol{x}$, $\boldsymbol{x}^{(m)}$; it holds that $\mathrm{tr}(J) = \mathbb{E}\|\boldsymbol{x}\|_2^2$ and $\mathrm{tr}(J_m) = \mathbb{E}\left\|\boldsymbol{x}^{(m)}\right\|_2^2$.

Finally, for $p \in [1, \infty]$ we use the standard notation p^* to denote the conjugate of p, that is, $p^* \in [1, \infty]$ and $\frac{1}{p} + \frac{1}{p^*} = 1$.

2. Global Rademacher Complexities in Multiple Kernel Learning

We first review global Rademacher complexities (GRC) in MKL. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be an i.i.d. sample drawn from P. The global Rademacher complexity is defined as

$$R(H_p) = \mathbb{E} \sup_{f_{\boldsymbol{w}} \in H_p} \langle \boldsymbol{w}, \frac{1}{n} \sum_{i=1}^n \sigma_i \boldsymbol{x}_i \rangle \tag{2}$$

where $(\sigma_i)_{1 \leq i \leq n}$ is an i.i.d. family (independent of $(\boldsymbol{x}_i)$) of Rademacher variables (random signs). Its empirical counterpart is denoted by $\widehat{R}(H_p) = \mathbb{E}\left[R(H_p) \middle| \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right] = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f_{\boldsymbol{w}} \in H_p} \langle \boldsymbol{w}, \frac{1}{n} \sum_{i=1}^n \sigma_i \boldsymbol{x}_i \rangle$. The interest in the global Rademacher complexity comes from that if known it can be used to bound the generalization error (Koltchinskii, 2001; Bartlett and Mendelson, 2002).

In the recent paper of Cortes et al. (2010) it was shown using a combinatorial argument that the empirical version of the global Rademacher complexity can be bounded as

$$\widehat{R}(H_p) \leq D \sqrt{\frac{cp^*}{2n} \left\| \left(\mathrm{tr}(K_m) \right)_{m=1}^M \right\|_{\frac{p^*}{2}}},$$

where $c = \frac{23}{22}$ and $\mathrm{tr}(K)$ denotes the trace of the kernel matrix K.

We will now show a quite short proof of this result, extending it to the whole range $p \in [1, \infty]$, but at the expense of a slightly worse constant, and then present a novel bound on the population version of the GRC. The proof presented here is based on the Khintchine-Kahane inequality (Kahane, 1985) using the constants taken from Lemma 3.3.1 and Proposition 3.4.1 in Kwapién and Woyczyński (1992).

Lemma 1 (Khintchine-Kahane inequality)**.** *Let be $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M \in \mathcal{H}$. Then, for any $q \geq 1$, it holds*

$$\mathbb{E}_{\boldsymbol{\sigma}} \Big\| \sum_{i=1}^{n} \sigma_i \boldsymbol{v}_i \Big\|_2^q \leq \Big(c \sum_{i=1}^{n} \big\| \boldsymbol{v}_i \big\|_2^2 \Big)^{\frac{q}{2}},$$

where $c = \max(1, q^* - 1)$. In particular the result holds for $c = q^*$.

Proposition 2 (Global Rademacher complexity, empirical version)**.** *For any $p \geq 1$ the empirical version of global Rademacher complexity of the multi-kernel class H_p can be bounded as*

$$\forall t \geq p: \quad \widehat{R}(H_p) \leq D \sqrt{\frac{t^*}{n} \Big\| \big(\operatorname{tr}(K_m) \big)_{m=1}^{M} \Big\|_{\frac{t^*}{2}}}.$$

Proof First note that it suffices to prove the result for $t = p$ as trivially $\|\boldsymbol{x}\|_{2,t} \leq \|\boldsymbol{x}\|_{2,p}$ holds for all $t \geq p$ and therefore $R(H_p) \leq R(H_t)$.

We can use a block-structured version of Hölder's inequality (cf. Lemma 15) and the Khintchine-Kahane (K.-K.) inequality (cf. Lemma 1) to bound the empirical version of the global Rademacher complexity as follows:

$$\begin{aligned}
\widehat{R}(H_p) &\overset{\text{def.}}{=} \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f_{\boldsymbol{w}} \in H_p} \langle \boldsymbol{w}, \frac{1}{n} \sum_{i=1}^{n} \sigma_i \boldsymbol{x}_i \rangle \\
&\overset{\text{Hölder}}{\leq} D \, \mathbb{E}_{\boldsymbol{\sigma}} \Big\| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \boldsymbol{x}_i \Big\|_{2,p^*} \\
&\overset{\text{Jensen}}{\leq} D \Big(\mathbb{E}_{\boldsymbol{\sigma}} \sum_{m=1}^{M} \Big\| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \boldsymbol{x}_i^{(m)} \Big\|_2^{p^*} \Big)^{\frac{1}{p^*}} \\
&\overset{\text{K.-K.}}{\leq} D \sqrt{\frac{p^*}{n}} \Big(\sum_{m=1}^{M} \Big(\underbrace{\frac{1}{n} \sum_{i=1}^{n} \big\| \boldsymbol{x}_i^{(m)} \big\|_2^2}_{= \operatorname{tr}(K_m)} \Big)^{\frac{p^*}{2}} \Big)^{\frac{1}{p^*}} \\
&= D \sqrt{\frac{p^*}{n} \Big\| \big(\operatorname{tr}(K_m) \big)_{m=1}^{M} \Big\|_{\frac{p^*}{2}}},
\end{aligned}$$

what was to show. ■

Note that there is a very good reason to state the above bound in terms of $t \geq p$ instead of solely in terms of p: the Rademacher complexity $\widehat{R}(H_p)$ is not monotonic in p and thus it is not always the best choice to take $t := p$ in the above bound. This can be readily seen, for example, for the easy case where all kernels have the same trace—in that case the bound translates into $\widehat{R}(H_p) \leq D \sqrt{t^* M^{\frac{2}{t^*}} \frac{\operatorname{tr}(K_1)}{n}}$. Interestingly, the function $x \mapsto x M^{2/x}$ is not monotone and attains its minimum for

$x = 2\log M$, where log denotes the natural logarithm with respect to the base e. This has interesting consequences: for any $p \le (2\log M)^*$ we can take the bound $\widehat{R}(H_p) \le D\sqrt{\frac{e\log(M)\operatorname{tr}(K_1)}{n}}$, which has only a mild dependency on the number of kernels; note that in particular we can take this bound for the ℓ_1-norm class $\widehat{R}(H_1)$ for all $M > 1$.

The above proof is very simple. However, computing the population version of the global Rademacher complexity of MKL is somewhat more involved and to the best of our knowledge has not been addressed yet by the literature. To this end, note that from the previous proof we obtain $R(H_p) \le \mathbb{E}\, D\sqrt{p^*/n}\big(\sum_{m=1}^M \big(\frac{1}{n}\sum_{i=1}^n \big\|\boldsymbol{x}_i^{(m)}\big\|_2^2\big)^{\frac{p^*}{2}}\big)^{\frac{1}{p^*}}$. We thus can use Jensen's inequality to move the expectation operator inside the root,

$$R(H_p) \le D\sqrt{p^*/n}\Big(\sum_{m=1}^M \mathbb{E}\big(\frac{1}{n}\sum_{i=1}^n \big\|\boldsymbol{x}_i^{(m)}\big\|_2^2\big)^{\frac{p^*}{2}}\Big)^{\frac{1}{p^*}}, \tag{3}$$

but now need a handle on the $\frac{p^*}{2}$-th moments. To this aim we use the inequalities of Rosenthal (1970) and Young (e.g., Steele, 2004) to show the following Lemma.

Lemma 3 (Rosenthal + Young)**.** *Let $X_1, \ldots, X_n$ be independent nonnegative random variables satisfying $\forall i : X_i \le B < \infty$ almost surely. Then, denoting $C_q = (2qe)^q$, for any $q \ge \frac{1}{2}$ it holds*

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^q \le C_q\left(\left(\frac{B}{n}\right)^q + \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i\right)^q\right).$$

The proof is defered to Appendix B. It is now easy to show:

Corollary 4 (Global Rademacher complexity, population version)**.** *Assume the kernels are uniformly bounded, that is, $\|k\|_\infty \le B < \infty$, almost surely. Then for any $p \ge 1$ the population version of global Rademacher complexity of the multi-kernel class H_p can be bounded as*

$$\forall t \ge p: \quad R(H_{p,D,M}) \le D\, t^* \sqrt{\frac{e}{n}\Big\|\big(\operatorname{tr}(J_m)\big)_{m=1}^M\Big\|_{\frac{t^*}{2}}} + \frac{\sqrt{Be}DM^{\frac{1}{t^*}}t^*}{n}.$$

For $t \ge 2$ the right-hand term can be discarded and the result also holds for unbounded kernels.

Proof As above in the previous proof it suffices to prove the result for $t = p$. From (3) we conclude by the previous Lemma

$$\begin{aligned} R(H_p) &\le D\sqrt{\frac{p^*}{n}}\left(\sum_{m=1}^M (ep^*)^{\frac{p^*}{2}}\left(\left(\frac{B}{n}\right)^{\frac{p^*}{2}} + \Big(\underbrace{\mathbb{E}\frac{1}{n}\sum_{i=1}^n \big\|\boldsymbol{x}_i^{(m)}\big\|_2^2}_{=\operatorname{tr}(J_m)}\Big)^{\frac{p^*}{2}}\right)\right)^{\frac{1}{p^*}} \\ &\le Dp^*\sqrt{\frac{e}{n}\Big\|\big(\operatorname{tr}(J_m)\big)_{m=1}^M\Big\|_{\frac{p^*}{2}}} + \frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n}, \end{aligned}$$

where for the last inequality we use the subadditivity of the root function. Note that for $p \ge 2$ it is $p^*/2 \le 1$ and thus it suffices to employ Jensen's inequality instead of the previous lemma so that we come along without the last term on the right-hand side. ■

For example, when the traces of the kernels are bounded, the above bound is essentially determined by $O\Big(\frac{p^* M^{\frac{1}{p^*}}}{\sqrt{n}}\Big)$. We can also remark that by setting $t = (\log(M))^*$ we obtain the bound $R(H_1) = O\Big(\frac{\log M}{\sqrt{n}}\Big)$.

2.1 Relation to Other Work

As discussed by Cortes et al. (2010), the above results lead to a generalization bound that improves on a previous result based on covering numbers by Srebro and Ben-David (2006). Another recently proposed approach to theoretically study MKL uses the Rademacher chaos complexity (RCC) (Ying and Campbell, 2009). The RCC is actually itself an upper bound on the usual Rademacher complexity. In their discussion, Cortes et al. (2010) observe that in the case $p = 1$ (traditional MKL), the bound of Proposition 2 grows logarithmically in the number of kernels M, and claim that the RCC approach would lead to a bound which is multiplicative in M. However, a closer look at the work of Ying and Campbell (2009) shows that this is not correct; in fact the RCC also leads to a logarithmic dependence in M when $p = 1$. This is because the RCC of a kernel class is the same as the RCC of its convex hull, and the RCC of the base class containing only the M individual kernels is logarithmic in M. This convex hull argument, however, only works for $p = 1$; we are unaware of any existing work trying to estimate the RCC or comparing it to the above approach in the case $p > 1$.

3. The Local Rademacher Complexity of Multiple Kernel Learning

We first give a gentle introduction to local Rademacher complexities in general and then present the main result of this paper: a lower and an upper bound on the local Rademacher complexity of ℓ_p-norm multiple kernel learning.

3.1 Local Rademacher Complexities in a Nutshell

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be an i.i.d. sample drawn from P; denote by $\mathbb{E}$ the expectation operator corresponding to P; let $\mathcal{F}$ be a class of functions mapping $\boldsymbol{x}_i$ to $\mathbb{R}$. Then the local Rademacher complexity is defined as

$$R_r(\mathcal{F}) = \mathbb{E} \sup_{f \in \mathcal{F} : Pf^2 \leq r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(\boldsymbol{x}_i)\,, \tag{4}$$

where $Pf^2 := \mathbb{E}(f(\boldsymbol{x}))^2$. In a nutshell, when comparing the global and local Rademacher complexities, that is, (2) and (4), we observe that the local one involves the additional constraint $Pf^2 \leq r$ on the (uncentered) "variance" of functions. It allows us to sort the functions according to their variances and discard the ones with suboptimal high variance. We can do so by, instead of McDiarmid's inequality, using more powerful concentration inequalities such as Talagrand's inequality (Talagrand, 1995). Roughly speaking, the local Rademacher complexity allows us to consider the problem at various scales simultaneously, leading to refined bounds. We will discuss this argument in more detail now. Our presentation is based on Koltchinskii (2006).

First, note that the classical (global) Rademacher theory of Bartlett and Mendelson (2002) and Koltchinskii (2001) gives an excess risk bound of the following form: $\exists C > 0$ so that with probability larger then $1 - \exp(-t)$ it holds

$$\left| P\hat{f} - Pf^* \right| \leq C\left(R(\mathcal{F}) + \sqrt{\frac{t}{n}} \right) =: \delta\,, \tag{5}$$

where $\hat{f} := \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(x_i)$, $f^* := \operatorname{argmin}_{f \in \mathcal{F}} Pf$, and $Pf := \mathbb{E} f(x)$. We denote the bound's value by δ and observe that, remarkably, if we consider the restricted class

$\mathcal{F}_\delta := \{f \in \mathcal{F} : |Pf - Pf^*| \leq \delta\}$, we have by (5) that $\hat{f} \in \mathcal{F}_\delta$ (and trivially $f^* \in \mathcal{F}_\delta$). This is remarkable and of significance because we can now state: with probability larger than $1 - \exp(-2t)$ it holds

$$\left|P\hat{f} - Pf^*\right| \leq C\left(R(\mathcal{F}_\delta) + \sqrt{\frac{t}{n}}\right). \tag{6}$$

The striking fact about the above inequality is that it depends on the complexity of the restricted class—no longer on the one of the original class; usually the complexity of the restricted class will be smaller than the one of the original class. Moreover, we can again denote the right-hand side of (6) by δ^{new} and repeat the argumentation. This way, we can step by step decrease the bound's value. If the bound (seen as a function in δ) defines a contraction, the limit of this iterative procedure is given by the fixed point of the bound.

This method has a serious limitation: although we can step by step decrease the Rademacher complexity occurring in the bound, the term $\sqrt{t/n}$ stays as it is and thus will hinder us from attaining a rate faster than $O(\sqrt{1/n})$. It would be desirable to have the term shrinking when passing to a smaller class $\mathcal{F}_\delta$. Can we replace the undesirable term by a more favorable one? And what properties would such a term need to have?

One of the basic foundations of learning theory are concentration inequalities (e.g., Bousquet et al., 2004). Even the most modern proof technique such as the fixed-point argument presented above can fail if it is built upon an insufficiently precise concentration inequality. As mentioned above, the stumbling block is the presence of the term $\sqrt{t/n}$ in the bound (5). The latter is a byproduct from the application of McDiarmid's inequality (McDiarmid, 1989)—a uniform version of Höffding's inequality—,which is used in Bartlett and Mendelson (2002) and Koltchinskii (2001) to relate the global Rademacher complexity with the excess risk.

The core idea now is that we can, instead of McDiarmid's inequality, use Talagrand's inequality (Talagrand, 1995), which is a uniform version of Bernstein's inequality. This gives

$$\left|P\hat{f} - Pf^*\right| \leq C\left(R(\mathcal{F}) + \sigma(\mathcal{F})\sqrt{\frac{t}{n}} + \frac{t}{n}\right) =: \delta. \tag{7}$$

Hereby $\sigma^2(\mathcal{F}) := \sup_{f\in\mathcal{F}} \mathbb{E}f^2$ is a bound on the (uncentered) "variance" of the functions considered. Now, denoting the right-hand side of (7) by δ, we obtain the following bound for the restricted class: $\exists C > 0$ so that with probability larger then $1 - \exp(-2t)$ it holds

$$\left|P\hat{f} - Pf^*\right| \leq C\left(R(\mathcal{F}_\delta) + \sigma(\mathcal{F}_\delta)\sqrt{\frac{t}{n}} + \frac{t}{n}\right). \tag{8}$$

As above, we denote the right-hand side of (8) by δ^{new} and repeat the argumentation. In general, we can expect the variance $\sigma^2(\mathcal{F}_\delta)$ to decrease step by step and if, seen as a function of δ, the bound defines a contraction, the limit is given by the fixed point of the bound.

It turns out that by this technique we can obtain fast convergence rates of the excess risk in the number of training examples n, which would be impossible by using global techniques such as the global Rademacher complexity or the Rademacher chaos complexity (Ying and Campbell, 2009), which—we recall—is in itself an upper bound on the global Rademacher complexity.

3.2 The Local Rademacher Complexity of MKL

In the context of ℓ_p-norm multiple kernel learning, we consider the hypothesis class H_p as defined in (1). Thus, given an i.i.d. sample $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ drawn from P, the local Rademacher complexity is given by $R_r(H_p) = \mathbb{E}\sup_{f_{\boldsymbol{w}}\in H_p: Pf_{\boldsymbol{w}}^2\le r}\langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^n \sigma_i \boldsymbol{x}_i\rangle$, where $Pf_{\boldsymbol{w}}^2 := \mathbb{E}(f_{\boldsymbol{w}}(\boldsymbol{x}))^2$.

We will need the following assumption for the case $1 \le p \le 2$:

Assumption (A) (low-correlation). *There exists a $c_\delta \in (0,1]$ such that, for any $m \neq m'$ and $\boldsymbol{w}_m \in \mathcal{H}_m, \boldsymbol{w}_{m'} \in \mathcal{H}_{m'}$, the Hilbert-space-valued variables $\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(M)}$ satisfy*

$$c_\delta \sum_{m=1}^M \mathbb{E}\left\langle \boldsymbol{w}_m, \boldsymbol{x}^{(m)}\right\rangle^2 \le \mathbb{E}\Big(\sum_{m=1}^M \left\langle \boldsymbol{w}_m, \boldsymbol{x}^{(m)}\right\rangle\Big)^2.$$

Since $\mathcal{H}_m, \mathcal{H}_{m'}$ are RKHSs with kernels $k_m, k_{m'}$, if we go back to the input random variable in the original space $X \in \mathcal{X}$, the above property means that for any fixed $t, t' \in \mathcal{X}$, the variables $k_m(X,t)$ and $k_{m'}(X,t')$ have a *low correlation*. In the most extreme case, $c_\delta = 1$, the variables are completely uncorrelated. This is the case, for example, if the original input space $\mathcal{X}$ is $\mathbb{R}^M$, the original input variable $X \in \mathcal{X}$ has independent coordinates, and the kernels $k_1,\ldots,k_M$ each act on a different coordinate. Such a setting was considered in particular by Raskutti et al. (2010) in the setting of ℓ_1-penalized MKL. We discuss this assumption in more detail in Section 6.3.

Note that, as self-adjoint, positive Hilbert-Schmidt operators, covariance operators enjoy discrete eigenvalue-eigenvector decompositions $J = \mathbb{E}\boldsymbol{x}\otimes\boldsymbol{x} = \sum_{j=1}^\infty \lambda_j \boldsymbol{u}_j \otimes \boldsymbol{u}_j$ and $J_m = \mathbb{E}\boldsymbol{x}^{(m)}\otimes\boldsymbol{x}^{(m)} = \sum_{j=1}^\infty \lambda_j^{(m)} \boldsymbol{u}_j^{(m)} \otimes \boldsymbol{u}_j^{(m)}$, where $(\boldsymbol{u}_j)_{j\ge1}$ and $(\boldsymbol{u}_j^{(m)})_{j\ge1}$ form orthonormal bases of $\mathcal{H}$ and $\mathcal{H}_m$, respectively.

We are now equipped to state our main results:

Theorem 5 (Local Rademacher complexity, $p \in [1,2]$)**.** *Assume that the kernels are uniformly bounded ($\|k\|_\infty \le B < \infty$) and that Assumption* **(A)** *holds. The local Rademacher complexity of the multi-kernel class H_p can be bounded for any $1 \le p \le 2$ as*

$$\forall t \in [p,2]: \quad R_r(H_p) \le \sqrt{\frac{16}{n}\left\|\left(\sum_{j=1}^\infty \min\left(rM^{1-\frac{2}{t^*}}, ceD^2t^{*2}\lambda_j^{(m)}\right)\right)_{m=1}^M\right\|_{\frac{t^*}{2}}} + \frac{\sqrt{Be}DM^{\frac{1}{t^*}}t^*}{n}.$$

Theorem 6 (Local Rademacher complexity, $p \ge 2$)**.** *The local Rademacher complexity of the multi-kernel class H_p can be bounded for any $p \ge 2$ as*

$$R_r(H_p) \le \sqrt{\frac{2}{n}\sum_{j=1}^\infty \min(r, D^2M^{\frac{2}{p^*}-1}\lambda_j)}.$$

Remark 7. *Note that for the case $p = 1$, by using $t = (\log(M))^*$ in Theorem 5, we obtain the bound*

$$R_r(H_1) \le \sqrt{\frac{16}{n}\left\|\left(\sum_{j=1}^\infty \min\left(rM, e^3D^2(\log M)^2\lambda_j^{(m)}\right)\right)_{m=1}^M\right\|_\infty} + \frac{\sqrt{B}e^{\frac{3}{2}}D\log(M)}{n},$$

for all $M \ge e^2$ (see below after the proof of Theorem 5 for a detailed justification).

Remark 8. *The result of Theorem 6 for $p \geq 2$ can be proved using considerably simpler techniques and without imposing assumptions on boundedness nor on uncorrelation of the kernels. If in addition the variables $(\boldsymbol{x}^{(m)})$ are centered and uncorrelated, then the spectra are related as follows :* $\operatorname{spec}(J) = \bigcup_{m=1}^{M} \operatorname{spec}(J_m)$*; that is,* $\{\lambda_i, i \geq 1\} = \bigcup_{m=1}^{M} \left\{\lambda_i^{(m)}, i \geq 1\right\}$*. Then one can write equivalently the bound of Theorem 6 as* $R_r(H_p) \leq \sqrt{\frac{2}{n}\sum_{m=1}^{M}\sum_{j=1}^{\infty}\min(r, D^2 M^{\frac{2}{p^*}-1}\lambda_j^{(m)})} = \sqrt{\frac{2}{n}\left\|\left(\sum_{j=1}^{\infty}\min(r, D^2 M^{\frac{2}{p^*}-1}\lambda_j^{(m)})\right)_{m=1}^{M}\right\|_1}$*. However, the main intended focus of this paper is on the more challenging case* $1 \leq p \leq 2$ *which is usually studied in multiple kernel learning and relevant in practice.*

Remark 9. *It is interesting to compare the above bounds for the special case $p = 2$ with the ones of Bartlett et al. (2005). The main term of the bound of Theorem 6 (taking $t = p = 2$) is then essentially determined by* $O\left(\sqrt{\frac{1}{n}\sum_{m=1}^{M}\sum_{j=1}^{\infty}\min\left(r, \lambda_j^{(m)}\right)}\right)$*. If the variables $(\boldsymbol{x}^{(m)})$ are centered and uncorrelated, by the relation between the spectra stated in Remark 8, this is equivalently of order* $O\left(\sqrt{\frac{1}{n}\sum_{j=1}^{\infty}\min\left(r, \lambda_j\right)}\right)$*, which is also what we obtain through Theorem 6, and coincides with the rate shown in Bartlett et al. (2005).*

Proof of Theorem 5 The proof is based on first relating the complexity of the class H_p with its centered counterpart, that is, where all functions $f_{\boldsymbol{w}} \in H_p$ are centered around their expected value. Then we compute the complexity of the centered class by decomposing the complexity into blocks, applying the no-correlation assumption, and using the inequalities of Hölder and Rosenthal. Then we relate it back to the original class, which we in the final step relate to a bound involving the truncation of the particular spectra of the kernels. Note that it suffices to prove the result for $t = p$ as trivially $R(H_p) \leq R(H_t)$ for all $p \leq t$.

STEP 1: RELATING THE ORIGINAL CLASS WITH THE CENTERED CLASS. In order to exploit the no-correlation assumption, we will work in large parts of the proof with the centered class $\tilde{H}_p = \left\{\tilde{f}_{\boldsymbol{w}} \;\middle|\; \|\boldsymbol{w}\|_{2,p} \leq D\right\}$, wherein $\tilde{f}_{\boldsymbol{w}} : \boldsymbol{x} \mapsto \langle \boldsymbol{w}, \tilde{\boldsymbol{x}}\rangle$, and $\tilde{\boldsymbol{x}} := \boldsymbol{x} - \mathbb{E}\boldsymbol{x}$. We start the proof by noting that $\tilde{f}_{\boldsymbol{w}}(\boldsymbol{x}) = f_{\boldsymbol{w}}(\boldsymbol{x}) - \langle \boldsymbol{w}, \mathbb{E}\boldsymbol{x}\rangle = f_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbb{E}\langle \boldsymbol{w}, \boldsymbol{x}\rangle = f_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbb{E}f_{\boldsymbol{w}}(\boldsymbol{x})$, so that, by the bias-variance decomposition, it holds that

$$Pf_{\boldsymbol{w}}^2 = \mathbb{E}f_{\boldsymbol{w}}(\boldsymbol{x})^2 = \mathbb{E}\left(f_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbb{E}f_w(\boldsymbol{x})\right)^2 + \left(\mathbb{E}f_w(\boldsymbol{x})\right)^2 = P\tilde{f}_{\boldsymbol{w}}^2 \;+\; \left(Pf_{\boldsymbol{w}}\right)^2. \tag{9}$$

Furthermore we note that by Jensen's inequality

$$\begin{aligned}\left\|\mathbb{E}\boldsymbol{x}\right\|_{2,p^*} &= \left(\sum_{m=1}^{M}\left\|\mathbb{E}\boldsymbol{x}^{(m)}\right\|_2^{p^*}\right)^{\frac{1}{p^*}} = \left(\sum_{m=1}^{M}\left\langle \mathbb{E}\boldsymbol{x}^{(m)}, \mathbb{E}\boldsymbol{x}^{(m)}\right\rangle^{\frac{p^*}{2}}\right)^{\frac{1}{p^*}} \\ &\overset{\text{Jensen}}{\leq} \left(\sum_{m=1}^{M}\mathbb{E}\left\langle \boldsymbol{x}^{(m)}, \boldsymbol{x}^{(m)}\right\rangle^{\frac{p^*}{2}}\right)^{\frac{1}{p^*}} \;=\; \sqrt{\left\|\left(\operatorname{tr}(J_m)\right)_{m=1}^{M}\right\|_{\frac{p^*}{2}}}\end{aligned} \tag{10}$$

so that we can express the complexity of the centered class in terms of the uncentered one as follows:

$$\begin{aligned} R_r(H_p) &= \mathbb{E} \sup_{\substack{f_{\boldsymbol{w}} \in H_p, \\ Pf_{\boldsymbol{w}}^2 \leq r}} \left\langle \boldsymbol{w}, \frac{1}{n} \sum_{i=1}^n \sigma_i \boldsymbol{x}_i \right\rangle \\ &\leq \mathbb{E} \sup_{\substack{f_{\boldsymbol{w}} \in H_p, \\ Pf_{\boldsymbol{w}}^2 \leq r}} \left\langle \boldsymbol{w}, \frac{1}{n} \sum_{i=1}^n \sigma_i \tilde{\boldsymbol{x}}_i \right\rangle + \mathbb{E} \sup_{\substack{f_{\boldsymbol{w}} \in H_p, \\ Pf_{\boldsymbol{w}}^2 \leq r}} \left\langle \boldsymbol{w}, \frac{1}{n} \sum_{i=1}^n \sigma_i \mathbb{E}\boldsymbol{x} \right\rangle \end{aligned}$$

Concerning the first term of the above upper bound, using (9) we have $P\tilde{f}_{\boldsymbol{w}}^2 \leq Pf_{\boldsymbol{w}}^2$, and thus

$$\mathbb{E} \sup_{\substack{f_{\boldsymbol{w}} \in H_p, \\ Pf_{\boldsymbol{w}}^2 \leq r}} \left\langle \boldsymbol{w}, \frac{1}{n} \sum_{i=1}^n \sigma_i \tilde{\boldsymbol{x}}_i \right\rangle \leq \mathbb{E} \sup_{\substack{f_{\boldsymbol{w}} \in H_p, \\ P\tilde{f}_{\boldsymbol{w}}^2 \leq r}} \left\langle \boldsymbol{w}, \frac{1}{n} \sum_{i=1}^n \sigma_i \tilde{\boldsymbol{x}}_i \right\rangle = R_r(\tilde{H}_p).$$

Now to bound the second term, we write

$$\begin{aligned} \mathbb{E} \sup_{\substack{f_{\boldsymbol{w}} \in H_p, \\ Pf_{\boldsymbol{w}}^2 \leq r}} \left\langle \boldsymbol{w}, \frac{1}{n} \sum_{i=1}^n \sigma_i \mathbb{E}\boldsymbol{x} \right\rangle &= \mathbb{E} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \right| \sup_{\substack{f_{\boldsymbol{w}} \in H_p, \\ Pf_{\boldsymbol{w}}^2 \leq r}} \langle \boldsymbol{w}, \mathbb{E}\boldsymbol{x} \rangle \\ &\leq \sup_{\substack{f_{\boldsymbol{w}} \in H_p, \\ Pf_{\boldsymbol{w}}^2 \leq r}} \langle \boldsymbol{w}, \mathbb{E}\boldsymbol{x} \rangle \left(\mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i \right)^2 \right)^{\frac{1}{2}} \\ &= \sqrt{n} \sup_{\substack{f_{\boldsymbol{w}} \in H_p, \\ Pf_{\boldsymbol{w}}^2 \leq r}} \langle \boldsymbol{w}, \mathbb{E}\boldsymbol{x} \rangle . \end{aligned}$$

Now observe finally that we have

$$\langle \boldsymbol{w}, \mathbb{E}\boldsymbol{x} \rangle \overset{\text{Hölder}}{\leq} \|\boldsymbol{w}\|_{2,p} \|\mathbb{E}\boldsymbol{x}\|_{2,p^*} \overset{(10)}{\leq} \|\boldsymbol{w}\|_{2,p} \sqrt{\left\| \left(\operatorname{tr}(J_m) \right)_{m=1}^M \right\|_{\frac{p^*}{2}}}$$

as well as

$$\langle \boldsymbol{w}, \mathbb{E}\boldsymbol{x} \rangle = \mathbb{E} f_{\boldsymbol{w}}(\boldsymbol{x}) \leq \sqrt{Pf_{\boldsymbol{w}}^2}.$$

We finally obtain, putting together the steps above,

$$R_r(H_p) \leq R_r(\tilde{H}_p) + n^{-\frac{1}{2}} \min \left(\sqrt{r}, D \sqrt{\left\| \left(\operatorname{tr}(J_m) \right)_{m=1}^M \right\|_{\frac{p^*}{2}}} \right) \tag{11}$$

This shows that we at the expense of the additional summand on the right hand side we can work with the centered class instead of the uncentered one.

STEP 2: BOUNDING THE COMPLEXITY OF THE CENTERED CLASS. Since the (centered) covariance operator $\mathbb{E}\tilde{\boldsymbol{x}}^{(m)} \otimes \tilde{\boldsymbol{x}}^{(m)}$ is also a self-adjoint Hilbert-Schmidt operator on $\mathcal{H}_m$, there exists an eigendecomposition

$$\mathbb{E}\tilde{\boldsymbol{x}}^{(m)} \otimes \tilde{\boldsymbol{x}}^{(m)} = \sum_{j=1}^{\infty} \tilde{\lambda}_j^{(m)} \tilde{\boldsymbol{u}}_j^{(m)} \otimes \tilde{\boldsymbol{u}}_j^{(m)}, \tag{12}$$

wherein $(\tilde{\boldsymbol{u}}_j^{(m)})_{j\geq 1}$ is an orthogonal basis of $\mathcal{H}_m$. Furthermore, the no-correlation assumption **(A)** entails $\mathbb{E}\tilde{\boldsymbol{x}}^{(l)}\otimes\tilde{\boldsymbol{x}}^{(m)}=\mathbf{0}$ for all $l\neq m$. As a consequence,

$$\begin{aligned} P\tilde{f}_{\boldsymbol{w}}^2 &= \mathbb{E}(f_{\boldsymbol{w}}(\tilde{\boldsymbol{x}}))^2 = \mathbb{E}\Big(\sum_{m=1}^M \big\langle \boldsymbol{w}_m, \tilde{\boldsymbol{x}}^{(m)}\big\rangle\Big)^2 = \sum_{l,m=1}^M \Big\langle \boldsymbol{w}_l, \big(\mathbb{E}\tilde{\boldsymbol{x}}^{(l)}\otimes\tilde{\boldsymbol{x}}^{(m)}\big)\boldsymbol{w}_m\Big\rangle \\ &\overset{\text{(A)}}{\geq} c_\delta \sum_{m=1}^M \Big\langle \boldsymbol{w}_m, \big(\mathbb{E}\tilde{\boldsymbol{x}}^{(m)}\otimes\tilde{\boldsymbol{x}}^{(m)}\big)\boldsymbol{w}_m\Big\rangle = \sum_{m=1}^M\sum_{j=1}^\infty \tilde{\lambda}_j^{(m)}\Big\langle \boldsymbol{w}_m, \tilde{\boldsymbol{u}}_j^{(m)}\Big\rangle^2 \end{aligned} \tag{13}$$

and, for all j and m,

$$\begin{aligned} \mathbb{E}\Big\langle \frac{1}{n}\sum_{i=1}^n \sigma_i\tilde{\boldsymbol{x}}_i^{(m)}, \tilde{\boldsymbol{u}}_j^{(m)}\Big\rangle^2 &= \mathbb{E}\frac{1}{n^2}\sum_{i,l=1}^n \sigma_i\sigma_l \Big\langle \tilde{\boldsymbol{x}}_i^{(m)}, \tilde{\boldsymbol{u}}_j^{(m)}\Big\rangle\Big\langle \tilde{\boldsymbol{x}}_l^{(m)}, \tilde{\boldsymbol{u}}_j^{(m)}\Big\rangle \overset{\sigma \text{ i.i.d.}}{=} \mathbb{E}\frac{1}{n^2}\sum_{i=1}^n \Big\langle \tilde{\boldsymbol{x}}_i^{(m)}, \tilde{\boldsymbol{u}}_j^{(m)}\Big\rangle^2 \\ &= \frac{1}{n}\Big\langle \tilde{\boldsymbol{u}}_j^{(m)}, \underbrace{\Big(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\tilde{\boldsymbol{x}}_i^{(m)}\otimes\tilde{\boldsymbol{x}}_i^{(m)}\Big)}_{=\mathbb{E}\tilde{\boldsymbol{x}}^{(m)}\otimes\tilde{\boldsymbol{x}}^{(m)}}\tilde{\boldsymbol{u}}_j^{(m)}\Big\rangle = \frac{\tilde{\lambda}_j^{(m)}}{n}. \end{aligned} \tag{14}$$

Now, let $h_1,\ldots,h_M$ be arbitrary nonnegative integers. We can express the local Rademacher complexity in terms of the eigendecomposition (12) as follows

$$\begin{aligned} R_r(\tilde{H}_p) &= \mathbb{E}\sup_{f_{\boldsymbol{w}}\in\tilde{H}_p: P\tilde{f}_{\boldsymbol{w}}^2\leq r}\Big\langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^n \sigma_i\tilde{\boldsymbol{x}}_i\Big\rangle \\ &= \mathbb{E}\sup_{f_{\boldsymbol{w}}\in\tilde{H}_p: P\tilde{f}_{\boldsymbol{w}}^2\leq r}\Big\langle \big(\boldsymbol{w}^{(m)}\big)_{m=1}^M, \big(\frac{1}{n}\sum_{i=1}^n \sigma_i\tilde{\boldsymbol{x}}_i^{(m)}\big)_{m=1}^M\Big\rangle \\ &= \mathbb{E}\sup_{f_{\boldsymbol{w}}\in\tilde{H}_p: P\tilde{f}_{\boldsymbol{w}}^2\leq r}\Big\langle \boldsymbol{w}, \Big(\sum_{j=1}^\infty \langle\frac{1}{n}\sum_{i=1}^n \sigma_i\tilde{\boldsymbol{x}}_i^{(m)}, \tilde{\boldsymbol{u}}_j^{(m)}\rangle\tilde{\boldsymbol{u}}_j^{(m)}\Big)_{m=1}^M\Big\rangle \\ &\overset{(\star)}{=} \mathbb{E}\sup_{P\tilde{f}_{\boldsymbol{w}}^2\leq r}\Big\langle \Big(\sum_{j=1}^{h_m}\sqrt{\tilde{\lambda}_j^{(m)}}\langle \boldsymbol{w}^{(m)}, \tilde{\boldsymbol{u}}_j^{(m)}\rangle\tilde{\boldsymbol{u}}_j^{(m)}\Big)_{m=1}^M, \\ &\qquad\qquad \Big(\sum_{j=1}^{h_m}\sqrt{\tilde{\lambda}_j^{(m)}}^{-1}\langle\frac{1}{n}\sum_{i=1}^n \sigma_i\tilde{\boldsymbol{x}}_i^{(m)}, \tilde{\boldsymbol{u}}_j^{(m)}\rangle\tilde{\boldsymbol{u}}_j^{(m)}\Big)_{m=1}^M\Big\rangle \\ &\quad + \mathbb{E}\sup_{f_{\boldsymbol{w}}\in\tilde{H}_p}\Big\langle \boldsymbol{w}, \Big(\sum_{j=h_m+1}^\infty \langle\frac{1}{n}\sum_{i=1}^n \sigma_i\tilde{\boldsymbol{x}}_i^{(m)}, \tilde{\boldsymbol{u}}_j^{(m)}\rangle\tilde{\boldsymbol{u}}_j^{(m)}\Big)_{m=1}^M\Big\rangle \\ &\overset{\text{C.-S., Jensen}}{\leq} \sup_{P\tilde{f}_{\boldsymbol{w}}^2\leq r}\Bigg[\Bigg(\sum_{m=1}^M\sum_{j=1}^{h_m}\tilde{\lambda}_j^{(m)}\langle \boldsymbol{w}^{(m)}, \tilde{\boldsymbol{u}}_j^{(m)}\rangle^2\Bigg)^{\frac{1}{2}} \\ &\qquad\qquad \times\Bigg(\sum_{m=1}^M\sum_{j=1}^{h_m}\Big(\tilde{\lambda}_j^{(m)}\Big)^{-1}\mathbb{E}\langle\frac{1}{n}\sum_{i=1}^n \sigma_i\tilde{\boldsymbol{x}}_i^{(m)}, \tilde{\boldsymbol{u}}_j^{(m)}\rangle^2\Bigg)^{\frac{1}{2}}\Bigg] \\ &\quad + \mathbb{E}\sup_{f_{\boldsymbol{w}}\in\tilde{H}_p}\Big\langle \boldsymbol{w}, \Big(\sum_{j=h_m+1}^\infty \langle\frac{1}{n}\sum_{i=1}^n \sigma_i\tilde{\boldsymbol{x}}_i^{(m)}, \tilde{\boldsymbol{u}}_j^{(m)}\rangle\tilde{\boldsymbol{u}}_j^{(m)}\Big)_{m=1}^M\Big\rangle, \end{aligned}$$

where for $(\star)$ we use the linearity of the scalar product, so that (13) and (14) yield

$$R_r(\tilde{H}_p) \overset{(13),\,(14)}{\leq} \sqrt{\frac{rc_\delta^{-1}\sum_{m=1}^M h_m}{n}} + \mathbb{E}\sup_{f_{\boldsymbol{w}}\in\tilde{H}_p}\left\langle \boldsymbol{w}, \Big(\sum_{j=h_m+1}^{\infty}\langle\frac{1}{n}\sum_{i=1}^n \sigma_i\tilde{\boldsymbol{x}}_i^{(m)},\tilde{\boldsymbol{u}}_j^{(m)}\rangle\tilde{\boldsymbol{u}}_j^{(m)}\Big)_{m=1}^M\right\rangle$$
$$\overset{\text{Hölder}}{\leq} \sqrt{\frac{rc_\delta^{-1}\sum_{m=1}^M h_m}{n}} + D\,\mathbb{E}\left\|\Big(\sum_{j=h_m+1}^{\infty}\langle\frac{1}{n}\sum_{i=1}^n \sigma_i\tilde{\boldsymbol{x}}_i^{(m)},\tilde{\boldsymbol{u}}_j^{(m)}\rangle\tilde{\boldsymbol{u}}_j^{(m)}\Big)_{m=1}^M\right\|_{2,p^*}.$$

STEP 3: KHINTCHINE-KAHANE'S AND ROSENTHAL'S INEQUALITIES. We can now use the Khintchine-Kahane (K.-K.) inequality (see Lemma 1 in Appendix B) to further bound the right term in the above expression as follows

$$\mathbb{E}\left\|\Big(\sum_{j=h_m+1}^{\infty}\langle\frac{1}{n}\sum_{i=1}^n \sigma_i\tilde{\boldsymbol{x}}_i^{(m)},\tilde{\boldsymbol{u}}_j^{(m)}\rangle\tilde{\boldsymbol{u}}_j^{(m)}\Big)_{m=1}^M\right\|_{2,p^*}$$
$$\overset{\text{Jensen}}{\leq} \mathbb{E}\Big(\sum_{m=1}^M \mathbb{E}_{\boldsymbol{\sigma}}\Big\|\sum_{j=h_m+1}^{\infty}\langle\frac{1}{n}\sum_{i=1}^n \sigma_i\tilde{\boldsymbol{x}}_i^{(m)},\tilde{\boldsymbol{u}}_j^{(m)}\rangle\tilde{\boldsymbol{u}}_j^{(m)}\Big\|_{\mathcal{H}_m}^{p^*}\Big)^{\frac{1}{p^*}}$$
$$\overset{\text{K.-K.}}{\leq} \sqrt{\frac{p^*}{n}}\,\mathbb{E}\Big(\sum_{m=1}^M\Big(\sum_{j=h_m+1}^{\infty}\frac{1}{n}\sum_{i=1}^n\langle\tilde{\boldsymbol{x}}_i^{(m)},\tilde{\boldsymbol{u}}_j^{(m)}\rangle^2\Big)^{\frac{p^*}{2}}\Big)^{\frac{1}{p^*}}$$
$$\overset{\text{Jensen}}{\leq} \sqrt{\frac{p^*}{n}}\,\Big(\sum_{m=1}^M\mathbb{E}\Big(\sum_{j=h_m+1}^{\infty}\frac{1}{n}\sum_{i=1}^n\langle\tilde{\boldsymbol{x}}_i^{(m)},\tilde{\boldsymbol{u}}_j^{(m)}\rangle^2\Big)^{\frac{p^*}{2}}\Big)^{\frac{1}{p^*}},$$

Note that for $p\geq 2$ it holds that $p^*/2\leq 1$, and thus it suffices to employ Jensen's inequality once again in order to move the expectation operator inside the inner term. In the general case we need a handle on the $\frac{p^*}{2}$-th moments and to this end employ Lemma 3 (Rosenthal + Young), which yields

$$\Big(\sum_{m=1}^M\mathbb{E}\Big(\sum_{j=h_m+1}^{\infty}\frac{1}{n}\sum_{i=1}^n\langle\tilde{\boldsymbol{x}}_i^{(m)},\tilde{\boldsymbol{u}}_j^{(m)}\rangle^2\Big)^{\frac{p^*}{2}}\Big)^{\frac{1}{p^*}}$$
$$\overset{\text{R+Y}}{\leq}\left(\sum_{m=1}^M(ep^*)^{\frac{p^*}{2}}\left(\Big(\frac{B}{n}\Big)^{\frac{p^*}{2}}+\Big(\sum_{j=h_m+1}^{\infty}\underbrace{\frac{1}{n}\sum_{i=1}^n\mathbb{E}\langle\tilde{\boldsymbol{x}}_i^{(m)},\tilde{\boldsymbol{u}}_j^{(m)}\rangle^2}_{=\tilde{\lambda}_j^{(m)}}\Big)^{\frac{p^*}{2}}\right)\right)^{\frac{1}{p^*}}$$
$$\overset{(*)}{\leq}\sqrt{ep^*\left(\frac{BM^{\frac{2}{p^*}}}{n}+\Big(\sum_{m=1}^M\Big(\sum_{j=h_m+1}^{\infty}\tilde{\lambda}_j^{(m)}\Big)^{\frac{p^*}{2}}\Big)^{\frac{2}{p^*}}\right)}$$
$$=\sqrt{ep^*\left(\frac{BM^{\frac{2}{p^*}}}{n}+\left\|\Big(\sum_{j=h_m+1}^{\infty}\tilde{\lambda}_j^{(m)}\Big)_{m=1}^M\right\|_{\frac{p^*}{2}}\right)}$$
$$\leq\sqrt{ep^*\left(\frac{BM^{\frac{2}{p^*}}}{n}+\left\|\Big(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\Big)_{m=1}^M\right\|_{\frac{p^*}{2}}\right)}$$

where for $(*)$ we used the subadditivity of $\sqrt[p^*]{\cdot}$ and in the last step we applied the Lidskii-Mirsky-Wielandt theorem which gives $\forall j,m:\ \tilde{\lambda}_j^{(m)} \leq \lambda_j^{(m)}$. Thus by the subadditivity of the root function

$$\begin{aligned} R_r(\tilde{H}_p) &\leq \sqrt{\frac{rc_\delta^{-1}\sum_{m=1}^M h_m}{n}} + D\sqrt{\frac{ep^{*2}}{n}\left(\frac{BM^{\frac{2}{p^*}}}{n} + \left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}\right)} \\ &= \sqrt{\frac{rc_\delta^{-1}\sum_{m=1}^M h_m}{n}} + \sqrt{\frac{ep^{*2}D^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}} + \frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n}. \end{aligned} \quad (15)$$

STEP 4: BOUNDING THE COMPLEXITY OF THE ORIGINAL CLASS. Now note that for all nonnegative integers h_m we either have

$$n^{-\frac{1}{2}}\min\left(\sqrt{rc_\delta^{-1}}, D\sqrt{\left\|\left(\mathrm{tr}(J_m)\right)_{m=1}^M\right\|_{\frac{p^*}{2}}}\right) \leq \sqrt{\frac{ep^{*2}D^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}}$$

(in case all h_m are zero) or it holds

$$n^{-\frac{1}{2}}\min\left(\sqrt{rc_\delta^{-1}}, D\sqrt{\left\|\left(\mathrm{tr}(J_m)\right)_{m=1}^M\right\|_{\frac{p^*}{2}}}\right) \leq \sqrt{\frac{rc_\delta^{-1}\sum_{m=1}^M h_m}{n}}$$

(in case that at least one h_m is nonzero) so that in any case we get

$$\begin{aligned} &n^{-\frac{1}{2}}\min\left(\sqrt{rc_\delta^{-1}}, D\sqrt{\left\|\left(\mathrm{tr}(J_m)\right)_{m=1}^M\right\|_{\frac{p^*}{2}}}\right) \\ &\qquad \leq \sqrt{\frac{rc_\delta^{-1}\sum_{m=1}^M h_m}{n}} + \sqrt{\frac{ep^{*2}D^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}}. \end{aligned} \quad (16)$$

Thus the following preliminary bound follows from (11) by (15) and (16):

$$R_r(H_p) \leq \sqrt{\frac{4rc_\delta^{-1}\sum_{m=1}^M h_m}{n}} + \sqrt{\frac{4ep^{*2}D^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}} + \frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n}, \quad (17)$$

for all nonnegative integers $h_m \geq 0$. We could stop here as the above bound is already the one that will be used in the subsequent section for the computation of the excess loss bounds. However, we can work a little more on the form of the bound to gain more insight on its properties—we will show that it is related to the truncation of the spectra at the scale r.

STEP 5: RELATING THE BOUND TO THE TRUNCATION OF THE SPECTRA OF THE KERNELS. To this end, notice that for all nonnegative real numbers A_1, A_2 and any $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}_+^m$ it holds for all $q \geq 1$

$$\sqrt{A_1} + \sqrt{A_2} \leq \sqrt{2(A_1+A_2)} \quad (18)$$

$$\|\boldsymbol{a}_1\|_q + \|\boldsymbol{a}_2\|_q \leq 2^{1-\frac{1}{q}}\|\boldsymbol{a}_1+\boldsymbol{a}_2\|_q \leq 2\|\boldsymbol{a}_1+\boldsymbol{a}_2\|_q \quad (19)$$

(the first statement follows from the concavity of the square root function and the second one is proved in appendix B; see Lemma 17) and thus

$$\begin{aligned}
&R_r(H_p)\\
&\overset{(18)}{\leq} \quad \sqrt{8\left(\frac{rc_\delta^{-1}\sum_{m=1}^M h_m}{n}+\frac{ep^{*2}D^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}\right)}+\frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n}\\
&\overset{\ell_1\text{-to-}\ell_{\frac{p^*}{2}}}{\leq} \quad \sqrt{\frac{8}{n}\left(rc_\delta^{-1}M^{1-\frac{2}{p^*}}\left\|\left(h_m\right)_{m=1}^M\right\|_{\frac{p^*}{2}}+ep^{*2}D^2\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}\right)}+\frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n}\\
&\overset{(19)}{\leq} \quad \sqrt{\frac{16}{n}\left\|\left(rc_\delta^{-1}M^{1-\frac{2}{p^*}}h_m+ep^{*2}D^2\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}}+\frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n},
\end{aligned}$$

where to obtain the second inequality we applied that for all non-negative $\boldsymbol{a}\in\mathbb{R}^M$ and $0<q<p\leq\infty$ it holds[2]

$$(\ell_q\text{-to-}\ell_p\text{ conversion})\qquad \|\boldsymbol{a}\|_q=\langle\mathbf{1},\boldsymbol{a}^q\rangle^{\frac{1}{q}}\overset{\text{Hölder}}{\leq}\left(\|\mathbf{1}\|_{(p/q)^*}\|\boldsymbol{a}^q\|_{p/q}\right)^{1/q}=M^{\frac{1}{q}-\frac{1}{p}}\|\boldsymbol{a}\|_p. \tag{20}$$

Since the above holds for all nonnegative integers h_m, it follows

$$\begin{aligned}
R_r(H_p) &\leq \sqrt{\frac{16}{n}\left\|\left(\min_{h_m\geq 0} rc_\delta^{-1}M^{1-\frac{2}{p^*}}h_m+ep^{*2}D^2\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}}+\frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n}\\
&= \sqrt{\frac{16}{n}\left\|\left(\sum_{j=1}^{\infty}\min\left(rc_\delta^{-1}M^{1-\frac{2}{p^*}},ep^{*2}D^2\lambda_j^{(m)}\right)\right)_{m=1}^M\right\|_{\frac{p^*}{2}}}+\frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n},
\end{aligned}$$

which completes the proof of the theorem. ■

Proof of Remark 7 To see that Remark 7 holds notice that $R(H_1)\leq R(H_p)$ for all $p\geq 1$ and thus by choosing $p=(\log(M))^*$ the above bound implies

$$\begin{aligned}
R_r(H_1) &\leq \sqrt{\frac{16}{n}\left\|\left(\sum_{j=1}^{\infty}\min\left(rc_\delta^{-1}M^{1-\frac{2}{p^*}},ep^{*2}D^2\lambda_j^{(m)}\right)\right)_{m=1}^M\right\|_{\frac{p^*}{2}}}+\frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n}\\
&\overset{\ell_{\frac{p^*}{2}}\text{-to-}\ell_\infty}{\leq} \sqrt{\frac{16}{n}\left\|\left(\sum_{j=1}^{\infty}\min\left(rc_\delta^{-1}M,ep^{*2}M^{\frac{2}{p^*}}D^2\lambda_j^{(m)}\right)\right)_{m=1}^M\right\|_{\infty}}+\frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n}\\
&= \sqrt{\frac{16}{n}\left\|\left(\sum_{j=1}^{\infty}\min\left(rc_\delta^{-1}M,e^3D^2(\log M)^2\lambda_j^{(m)}\right)\right)_{m=1}^M\right\|_{\infty}}+\frac{\sqrt{Be}^{\frac{3}{2}}D(\log M)}{n},
\end{aligned}$$

which completes the proof. ■

2. We denote by $\boldsymbol{a}^q$ the vector with entries a_i^q and by $\mathbf{1}$ the vector with entries all 1.

Proof of Theorem 6.

The eigendecomposition $\mathbb{E}\boldsymbol{x}\otimes\boldsymbol{x}=\sum_{j=1}^{\infty}\lambda_j\boldsymbol{u}_j\otimes\boldsymbol{u}_j$ yields

$$Pf_{\boldsymbol{w}}^2=\mathbb{E}(f_{\boldsymbol{w}}(\boldsymbol{x}))^2=\mathbb{E}\langle\boldsymbol{w},\boldsymbol{x}\rangle^2=\big\langle\boldsymbol{w},(\mathbb{E}\boldsymbol{x}\otimes\boldsymbol{x})\boldsymbol{w}\big\rangle=\sum_{j=1}^{\infty}\lambda_j\big\langle\boldsymbol{w},\boldsymbol{u}_j\big\rangle^2, \tag{21}$$

and, for all j

$$\begin{aligned}\mathbb{E}\Big\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{x}_i,\boldsymbol{u}_j\Big\rangle^2 &= \mathbb{E}\frac{1}{n^2}\sum_{i,l=1}^{n}\sigma_i\sigma_l\big\langle\boldsymbol{x}_i,\boldsymbol{u}_j\big\rangle\big\langle\boldsymbol{x}_l,\boldsymbol{u}_j\big\rangle \overset{\sigma\text{ i.i.d.}}{=}\mathbb{E}\frac{1}{n^2}\sum_{i=1}^{n}\big\langle\boldsymbol{x}_i,\boldsymbol{u}_j\big\rangle^2\\ &= \frac{1}{n}\Big\langle\boldsymbol{u}_j,\Big(\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\boldsymbol{x}_i\otimes\boldsymbol{x}_i}_{=\mathbb{E}\boldsymbol{x}\otimes\boldsymbol{x}}\Big)\boldsymbol{u}_j\Big\rangle=\frac{\lambda_j}{n}.\end{aligned} \tag{22}$$

Therefore, we can use, for any nonnegative integer h, the Cauchy-Schwarz inequality and a block-structured version of Hölder's inequality (see Lemma 15) to bound the local Rademacher complexity as follows:

$$\begin{aligned}R_r(H_p) &= \mathbb{E}\sup_{f_{\boldsymbol{w}}\in H_p:Pf_{\boldsymbol{w}}^2\le r}\Big\langle\boldsymbol{w},\frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{x}_i\Big\rangle\\ &= \mathbb{E}\sup_{f_{\boldsymbol{w}}\in H_p:Pf_{\boldsymbol{w}}^2\le r}\Big\langle\sum_{j=1}^{h}\sqrt{\lambda_j}\langle\boldsymbol{w},\boldsymbol{u}_j\rangle\boldsymbol{u}_j,\sum_{j=1}^{h}\sqrt{\lambda_j}^{-1}\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{x}_i,\boldsymbol{u}_j\rangle\boldsymbol{u}_j\Big\rangle\\ &\qquad+\Big\langle\boldsymbol{w},\sum_{j=h+1}^{\infty}\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{x}_i,\boldsymbol{u}_j\rangle\boldsymbol{u}_j\Big\rangle\\ &\overset{\text{C.-S., (21), (22)}}{\le}\sqrt{\frac{rh}{n}}+\mathbb{E}\sup_{f_{\boldsymbol{w}}\in H_p}\Big\langle\boldsymbol{w},\sum_{j=h+1}^{\infty}\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{x}_i,\boldsymbol{u}_j\rangle\boldsymbol{u}_j\Big\rangle\\ &\overset{\text{Hölder}}{\le}\sqrt{\frac{rh}{n}}+D\mathbb{E}\Big\|\sum_{j=h+1}^{\infty}\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{x}_i,\boldsymbol{u}_j\rangle\boldsymbol{u}_j\Big\|_{2,p^*}\\ &\overset{\ell_{\frac{p^*}{2}}\text{-to-}\ell_2}{\le}\sqrt{\frac{rh}{n}}+DM^{\frac{1}{p^*}-\frac{1}{2}}\mathbb{E}\Big\|\sum_{j=h+1}^{\infty}\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{x}_i,\boldsymbol{u}_j\rangle\boldsymbol{u}_j\Big\|_{\mathcal{H}}\\ &\overset{\text{Jensen}}{\le}\sqrt{\frac{rh}{n}}+DM^{\frac{1}{p^*}-\frac{1}{2}}\Bigg(\sum_{j=h+1}^{\infty}\underbrace{\mathbb{E}\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{x}_i,\boldsymbol{u}_j\rangle^2}_{\overset{(22)}{\le}\frac{\lambda_j}{n}}\Bigg)^{\frac{1}{2}}\\ &\le\sqrt{\frac{rh}{n}}+\sqrt{\frac{D^2M^{\frac{2}{p^*}-1}}{n}\sum_{j=h+1}^{\infty}\lambda_j}.\end{aligned}$$

Since the above holds for all h, the result now follows from $\sqrt{A}+\sqrt{B}\le\sqrt{2(A+B)}$ for all nonnegative real numbers A,B (which holds by the concavity of the square root function):

$$R_r(H_p)\le\sqrt{\frac{2}{n}\min_{0\le h\le n}\Big(rh+D^2M^{\frac{2}{p^*}-1}\sum_{j=h+1}^{\infty}\lambda_j\Big)}=\sqrt{\frac{2}{n}\sum_{j=1}^{\infty}\min(r,D^2M^{\frac{2}{p^*}-1}\lambda_j)}.$$

4. Lower Bound

In this subsection we investigate the tightness of our bound on the local Rademacher complexity of H_p. To derive a lower bound we consider the particular case where variables $\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(M)}$ are i.i.d. For example, this happens if the original input space $\mathcal{X}$ is $\mathbb{R}^M$, the original input variable $X \in \mathcal{X}$ has i.i.d. coordinates, and the kernels $k_1,\ldots,k_M$ are identical and each act on a different coordinate of X.

Lemma 10. *Assume that the variables $\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(M)}$ are centered and identically independently distributed. Then, the following lower bound holds for the local Rademacher complexity of H_p for any $p \geq 1$:*

$$R_r(H_{p,D,M}) \;\geq\; R_{rM}(H_{1,DM^{1/p^*},1}).$$

Proof First note that since the $\boldsymbol{x}^{(i)}$ are centered and uncorrelated, that

$$Pf_{\boldsymbol{w}}^2 = \Big(\sum_{m=1}^{M} \big\langle \boldsymbol{w}_m, \boldsymbol{x}^{(m)} \big\rangle\Big)^2 = \sum_{m=1}^{M} \big\langle \boldsymbol{w}_m, \boldsymbol{x}^{(m)} \big\rangle^2.$$

Now it follows

$$\begin{aligned}
R_r(H_{p,D,M}) &= \mathbb{E} \sup_{\substack{\boldsymbol{w}\colon \\ \substack{Pf_{\boldsymbol{w}}^2 \leq r \\ \|\boldsymbol{w}\|_{2,p} \leq D}}} \Big\langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^{n} \sigma_i \boldsymbol{x}_i \Big\rangle \\
&= \mathbb{E} \sup_{\substack{\boldsymbol{w}\colon \\ \substack{\sum_{m=1}^{M} \langle \boldsymbol{w}^{(m)}, \boldsymbol{x}^{(m)} \rangle^2 \leq r \\ \|\boldsymbol{w}\|_{2,p} \leq D}}} \Big\langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^{n} \sigma_i \boldsymbol{x}_i \Big\rangle \\
&\geq \mathbb{E} \sup_{\substack{\boldsymbol{w}\colon \\ \substack{\forall m\colon \langle \boldsymbol{w}^{(m)}, \boldsymbol{x}^{(m)} \rangle^2 \leq r/M \\ \left\|\boldsymbol{w}^{(m)}\right\|_{2,p} \leq D \\ \left\|\boldsymbol{w}^{(1)}\right\| = \cdots = \left\|\boldsymbol{w}^{(M)}\right\|}}} \Big\langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^{n} \sigma_i \boldsymbol{x}_i \Big\rangle \\
&= \mathbb{E} \sup_{\substack{\boldsymbol{w}\colon \\ \substack{\forall m\colon \langle \boldsymbol{w}^{(m)}, \boldsymbol{x}^{(m)} \rangle^2 \leq r/M \\ \forall m\colon \left\|\boldsymbol{w}^{(m)}\right\|_2 \leq DM^{-\frac{1}{p}}}}} \sum_{m=1}^{M} \Big\langle \boldsymbol{w}^{(m)}, \frac{1}{n}\sum_{i=1}^{n} \sigma_i \boldsymbol{x}_i^{(m)} \Big\rangle \\
&= \sum_{m=1}^{M} \mathbb{E} \sup_{\substack{\boldsymbol{w}^{(m)}\colon \\ \substack{\langle \boldsymbol{w}^{(m)}, \boldsymbol{x}^{(m)} \rangle^2 \leq r/M \\ \left\|\boldsymbol{w}^{(m)}\right\|_2 \leq DM^{-\frac{1}{p}}}}} \Big\langle \boldsymbol{w}^{(m)}, \frac{1}{n}\sum_{i=1}^{n} \sigma_i \boldsymbol{x}_i^{(m)} \Big\rangle \;,
\end{aligned}$$

so that we can use the i.i.d. assumption on $\boldsymbol{x}^{(m)}$ to equivalently rewrite the last term as follows:

$$\begin{aligned} R_r(H_{p,D,M}) &\overset{\boldsymbol{x}^{(m)}\text{ i.i.d.}}{\geq} \mathbb{E} \sup_{\boldsymbol{w}^{(1)}:\ \substack{\langle \boldsymbol{w}^{(1)}, \boldsymbol{x}^{(1)}\rangle^2 \leq r/M \\ \left\|\boldsymbol{w}^{(1)}\right\|_2 \leq DM^{-\frac{1}{p}}}} \left\langle M\boldsymbol{w}^{(1)}, \frac{1}{n}\sum_{i=1}^n \sigma_i \boldsymbol{x}_i^{(1)} \right\rangle \\ &= \mathbb{E} \sup_{\boldsymbol{w}^{(1)}:\ \substack{\langle M\boldsymbol{w}^{(1)}, \boldsymbol{x}^{(1)}\rangle^2 \leq rM \\ \left\|M\boldsymbol{w}^{(1)}\right\|_2 \leq DM^{\frac{1}{p^*}}}} \left\langle M\boldsymbol{w}^{(1)}, \frac{1}{n}\sum_{i=1}^n \sigma_i \boldsymbol{x}_i^{(1)} \right\rangle \\ &= \mathbb{E} \sup_{\boldsymbol{w}^{(1)}:\ \substack{\langle \boldsymbol{w}^{(1)}, \boldsymbol{x}^{(1)}\rangle^2 \leq rM \\ \left\|\boldsymbol{w}^{(1)}\right\|_2 \leq DM^{\frac{1}{p^*}}}} \left\langle \boldsymbol{w}^{(1)}, \frac{1}{n}\sum_{i=1}^n \sigma_i \boldsymbol{x}_i^{(1)} \right\rangle \\ &= R_{rM}(H_{1,DM^{1/p^*},1}) \end{aligned}$$

■

In Mendelson (2003) it was shown that there is an absolute constant c so that if $\lambda^{(1)} \geq \frac{1}{n}$ then for all $r \geq \frac{1}{n}$ it holds $R_r(H_{1,1,1}) \geq \sqrt{\frac{c}{n}\sum_{j=1}^\infty \min(r, \lambda_j^{(1)})}$. Closer inspection of the proof reveals that more generally it holds $R_r(H_{1,D,1}) \geq \sqrt{\frac{c}{n}\sum_{j=1}^\infty \min(r, D^2\lambda_j^{(1)})}$ if $\lambda_1^{(m)} \geq \frac{1}{nD^2}$ so that we can use that result together with the previous lemma to obtain:

Theorem 11 (Lower bound)**.** *Assume that the kernels are centered and identically independently distributed. Then, the following lower bound holds for the local Rademacher complexity of H_p. There is an absolute constant c such that if $\lambda^{(1)} \geq \frac{1}{nD^2}$ then for all $r \geq \frac{1}{n}$ and $p \geq 1$,*

$$R_r(H_{p,D,M}) \ \geq\ \sqrt{\frac{c}{n}\sum_{j=1}^\infty \min(rM, D^2M^{2/p^*}\lambda_j^{(1)})}. \tag{23}$$

We would like to compare the above lower bound with the upper bound of Theorem 5. To this end note that for centered identical independent kernels the upper bound reads

$$R_r(H_p) \leq \sqrt{\frac{16}{n}\sum_{j=1}^\infty \min\left(rM, ceD^2p^{*2}M^{\frac{2}{p^*}}\lambda_j^{(1)}\right)} + \frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n},$$

which is of the order $O\big(\sqrt{\sum_{j=1}^\infty \min\left(rM, D^2M^{\frac{2}{p^*}}\lambda_j^{(1)}\right)}\big)$ and, disregarding the quickly converging term on the right hand side and absolute constants, again matches the upper bounds of the previous section. A similar comparison can be performed for the upper bound of Theorem 6: by Remark 8 the bound reads

$$R_r(H_p) \leq \sqrt{\frac{2}{n}\left\|\Big(\sum_{j=1}^\infty \min(r, D^2M^{\frac{2}{p^*}-1}\lambda_j^{(m)})\Big)_{m=1}^M\right\|_1},$$

which for i.i.d. kernels becomes $\sqrt{2/n\sum_{j=1}^\infty \min\left(rM, D^2M^{\frac{2}{p^*}}\lambda_j^{(1)}\right)}$ and thus, besides absolute constants, matches the lower bound. This shows that the upper bounds of the previous section are tight.

5. Excess Risk Bounds

In this section we show an application of our results to prediction problems, such as classification or regression. To this aim, in addition to the data $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ introduced earlier in this paper, let also a label sequence $y_1,\ldots,y_n \subset [-1,1]$ be given that is i.i.d. generated from a probability distribution. The goal in statistical learning is to find a hypothesis f from a pregiven class $\mathcal{F}$ that minimizes the expected loss $\mathbb{E}\, l(f(\boldsymbol{x}),y)$, where $l : \mathbb{R}^2 \mapsto [-1,1]$ is a predefined loss function that encodes the objective of the given learning/prediction task at hand. For example, the hinge loss $l(t,y) = \max(0,1-yt)$ and the squared loss $l(t,y) = (t-y)^2$ are frequently used in classification and regression problems, respectively.

Since the distribution generating the example/label pairs is unknown, the optimal decision function

$$f^* := \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}\, l(f(\boldsymbol{x}),y)$$

can not be computed directly and a frequently used method consists of instead minimizing the *empirical* loss,

$$\hat{f} := \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} l(f(\boldsymbol{x}_i),y_i).$$

In order to evaluate the performance of this so-called *empirical risk minimization* (ERM) algorithm we study the excess loss,

$$P(l_{\hat{f}} - l_{f^*}) \;:=\; \mathbb{E}\, l(\hat{f}(\boldsymbol{x}),y) - \mathbb{E}\, l(f^*(\boldsymbol{x}),y).$$

In Bartlett et al. (2005) and Koltchinskii (2006) it was shown that the rate of convergence of the excess risk is basically determined by the fixed point of the local Rademacher complexity. For example, the following result is a slight modification of Corollary 5.3 in Bartlett et al. (2005) that is well-tailored to the class studied in this paper.[3]

Lemma 12. *Let $\mathcal{F}$ be an absolute convex class ranging in the interval $[a,b]$ and let l be a Lipschitz continuous loss with constant L. Assume there is a positive constant F such that*

$$\forall f \in \mathcal{F} :\; P(f-f^*)^2 \le F\, P(l_f - l_{f^*}). \tag{24}$$

Then, denoting by r^* the fixed point of

$$2FL\, R_{\frac{r}{4L^2}}(\mathcal{F})$$

for all $x > 0$ with probability at least $1-e^{-x}$ the excess loss can be bounded as

$$P(l_{\hat{f}} - l_{f^*}) \le 7\frac{r^*}{F} + \frac{(11L(b-a)+27F)x}{n}.$$

Note that condition (24) on the loss function is fulfilled, for example, when the kernel is uniformly bounded and the loss function is strongly convex and Lipschitz continuous on the domain considered (Bartlett et al., 2006). This includes, for example, the squared loss as defined above, the

3. We exploit the improved constants from Theorem 3.3 in Bartlett et al. (2005) because an absolute convex class is star-shaped. Compared to Corollary 5.3 in Bartlett et al. (2005) we also use a slightly more general function class ranging in $[a,b]$ instead of the interval $[-1,1]$. This is also justified by Theorem 3.3.

logistic loss $l(t,y) = \ln(1+\exp(-yt))$, and the exponential loss $l(t,y) = \exp(-yt)$. The case of the hinge loss (see definition above) is more delicate, since it is not a strongly convex loss function. In general, the hinge loss does not satisfy (24) on an arbitrary convex class $\mathcal{F}$; for this reason, there is no direct, general "fast rate" excess loss analogue to the popular margin-radius bounds obtained through global Rademacher analysis. Nevertheless, local Rademacher complexity analysis can still be put to good use for algorithms based on the hinge loss. In fact, the hinge loss satisfies, under an additional "noise exponent condition" assumption, a restricted version of (24), namely, when f^* is taken equal to the Bayes classifier. This can be used to study theoretically the behavior of penalized ERM methods such as the support vector machine, and more precisely to obtain oracle-type inequalities (this roughly means that the penalized ERM can be shown to pick a correct trade-off of bias and estimation error, leading to fast convergence rates). In this sense, the local Rademacher complexity bound we have presented here can in principle be plugged in into the SVM analysis of Blanchard et al. (2008), directly replacing the local Rademacher analysis for a single kernel studied there under setting (S1); see also Steinwart and Christmann (2008, Chapter 8) for a comparable analysis. This more elaborate analysis does, however, not fall directly into the scope of the comparably simpler result of Lemma 12, which considers simple ERM over a fixed model, so that we refer the reader to the references cited above for more details.

Lemma 12 shows that in order to obtain an excess risk bound on the multi-kernel class H_p it suffices to compute the fixed point of our bound on the local Rademacher complexity presented in Section 3. To this end we show:

Lemma 13. *Assume that $\|k\|_\infty \le B$ almost surely and assumption (A) holds; let $p \in [1,2]$. For the fixed point r^* of the local Rademacher complexity $2FLR_{\frac{r}{4L^2}}(H_p)$ it holds*

$$r^* \le \min_{0\le h_m\le\infty} \frac{4c_\delta^{-1}F^2\sum_{m=1}^M h_m}{n} + 8FL\sqrt{\frac{ep^{*2}D^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}} + \frac{4\sqrt{Be}DFLM^{\frac{1}{p^*}}p^*}{n}\,.$$

Proof For this proof we make use of the bound (17) on the local Rademacher complexity. Defining

$$a = \frac{4c_\delta^{-1}F^2\sum_{m=1}^M h_m}{n} \quad \text{and} \quad b = 4FL\sqrt{\frac{ep^{*2}D^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}} + \frac{2\sqrt{Be}DFLM^{\frac{1}{p^*}}p^*}{n}\,,$$

in order to find a fixed point of (17) we need to solve for $r = \sqrt{ar}+b$, which is equivalent to solving $r^2-(a+2b)r+b^2=0$ for a positive root. Denote this solution by r^*. It is then easy to see that $r^* \ge a+2b$. Resubstituting the definitions of a and b yields the result. ■

We now address the issue of computing actual rates of convergence of the fixed point r^* under the assumption of algebraically decreasing eigenvalues of the kernel matrices, this means, we assume $\exists d_m:\ \lambda_j^{(m)} \le d_m j^{-\alpha_m}$ for some $\alpha_m > 1$. This is a common assumption and, for example, met for finite rank kernels and convolution kernels (Williamson et al., 2001). Notice that this implies

$$\begin{aligned}\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)} &\le d_m\sum_{j=h_m+1}^{\infty} j^{-\alpha_m} \le d_m\int_{h_m}^{\infty}x^{-\alpha_m}dx = d_m\left[\frac{1}{1-\alpha_m}x^{1-\alpha_m}\right]_{h_m}^{\infty}\\ &= -\frac{d_m}{1-\alpha_m}h_m^{1-\alpha_m}\,. \qquad (25)\end{aligned}$$

To exploit the above fact, first note that by ℓ_p-to-ℓ_q conversion

$$\frac{4c_\delta^{-1}F^2\sum_{m=1}^M h_m}{n} \;\le\; 4F\sqrt{\frac{c_\delta^{-1}F^2M\sum_{m=1}^M h_m^2}{n^2}} \;\le\; 4F\sqrt{\frac{c_\delta^{-1}F^2M^{2-\frac{2}{p^*}}\left\|\left(h_m^2\right)\right)_{m=1}^M\right\|_{2/p^*}}{n^2}}$$

so that we can translate the result of the previous lemma by (18), (19), and (20) into

$$\begin{aligned} r^* \;\le\; \min_{0\le h_m\le\infty} 8F\sqrt{\frac{1}{n}\left\|\left(\frac{c_\delta^{-1}F^2M^{2-\frac{2}{p^*}}h_m^2}{n}+4ep^{*2}D^2L^2\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}} \\ +\frac{4\sqrt{Be}DFLM^{\frac{1}{p^*}}p^*}{n}\,. \end{aligned} \tag{26}$$

Inserting the result of (25) into the above bound and setting the derivative with respect to h_m to zero we find the optimal h_m as

$$h_m = \left(4c_\delta d_m e p^{*2}D^2F^{-2}L^2M^{\frac{2}{p^*}-2}n\right)^{\frac{1}{1+\alpha_m}}.$$

Resubstituting the above into (26) we note that

$$r^* = O\left(\sqrt{\left\|\left(n^{-\frac{2\alpha_m}{1+\alpha_m}}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}}\right)$$

so that we observe that the asymptotic rate of convergence in n is determined by the kernel with the smallest decreasing spectrum (i.e., smallest α_m). Denoting $d_{\max} := \max_{m=1,\ldots,M} d_m$, $\alpha_{\min} := \min_{m=1,\ldots,M}\alpha_m$, and $h_{\max} := \left(4c_\delta d_{\max} e p^{*2}D^2F^{-2}L^2M^{\frac{2}{p^*}-2}n\right)^{\frac{1}{1+\alpha_{\min}}}$ we can upper-bound (26) by

$$\begin{aligned} r^* &\le 8F\sqrt{\frac{3-\alpha_{\min}}{1-\alpha_{\min}}c_\delta^{-1}F^2M^2h_{\max}^2n^{-2}}+\frac{4\sqrt{Be}DFLM^{\frac{1}{p^*}}p^*}{n} \\ &\le 8\sqrt{\frac{3-\alpha_{\min}}{1-\alpha_{\min}}}c_\delta^{-1}F^2Mh_{\max}n^{-1}+\frac{4\sqrt{Be}DFLM^{\frac{1}{p^*}}p^*}{n} \\ &\le 16\sqrt{e\frac{3-\alpha_{\min}}{1-\alpha_{\min}}}c_\delta^{-1}(d_{\max}D^2L^2p^{*2})^{\frac{1}{1+\alpha_{\min}}}F^{\frac{2\alpha_{\min}}{1+\alpha_{\min}}}M^{1+\frac{2}{1+\alpha_{\min}}\left(\frac{1}{p^*}-1\right)}n^{-\frac{\alpha_{\min}}{1+\alpha_{\min}}} \\ &\quad+\frac{4\sqrt{Be}DFLM^{\frac{1}{p^*}}p^*}{n}\,. \end{aligned} \tag{27}$$

We have thus proved the following theorem, which follows by the above inequality, Lemma 12, and the fact that our class H_p ranges in $BDM^{\frac{1}{p^*}}$.

Theorem 14. *Assume that $\|k\|_\infty \le B$, assumption (A) holds, and it $\exists d_{\max} > 0$ and $\alpha := \alpha_{\min} > 1$ such that for all $m = 1,\ldots,M$ it holds $\lambda_j^{(m)} \le d_{\max} j^{-\alpha}$. Let l be a Lipschitz continuous loss with constant L and assume there is a positive constant F such that $\forall f \in \mathcal{F}:\ P(f-f^*)^2 \le F\,P(l_f - l_{f^*})$.*

Then for all $x > 0$ with probability at least $1 - e^{-x}$ the excess loss of the multi-kernel class H_p can be bounded for $p \in [1,\ldots,2]$ as

$$P(l_{\hat{f}} - l_{f^*}) \leq \min_{t\in[p,2]} 186\sqrt{\frac{3-\alpha}{1-\alpha}}\, c_\delta^{\frac{1-\alpha}{1+\alpha}} \left(d_{\max} D^2 L^2 t^{*2}\right)^{\frac{1}{1+\alpha}} F^{\frac{\alpha-1}{\alpha+1}} M^{1+\frac{2}{1+\alpha}\left(\frac{1}{t^*}-1\right)} n^{-\frac{\alpha}{1+\alpha}} + \frac{47\sqrt{B}DLM^{\frac{1}{t^*}}t^*}{n} + \frac{(22BDLM^{\frac{1}{t^*}} + 27F)x}{n}$$

We see from the above bound that convergence can be almost as slow as $O\left(p^* M^{\frac{1}{p^*}} n^{-\frac{1}{2}}\right)$ (if at least one $\alpha_m \approx 1$ is small and thus $\alpha_{\min}$ is small) and almost as fast as $O\left(n^{-1}\right)$ (if α_m is large for all m and thus $\alpha_{\min}$ is large). For example, the latter is the case if all kernels have finite rank and also the convolution kernel is an example of this type.

Notice that we of course could repeat the above discussion to obtain excess risk bounds for the case $p \geq 2$ as well, but since it is very questionable that this will lead to new insights, it is omitted for simplicity.

6. Discussion

In this section we compare the obtained local Rademacher bound with the global one, discuss related work as well as the assumption **(A)**, and give a practical application of the bounds by studying the appropriateness of small/large p in various learning scenarios.

6.1 Global vs. Local Rademacher Bounds

In this section, we discuss the rates obtained from the bound in Theorem 14 for the excess risk and compare them to the rates obtained using the global Rademacher complexity bound of Corollary 4. To simplify somewhat the discussion, we assume that the eigenvalues satisfy $\lambda_j^{(m)} \leq d\, j^{-\alpha}$ (with $\alpha > 1$) for all m and concentrate on the rates obtained as a function of the parameters n, α, M, D and p, while considering other parameters fixed and hiding them in a big-O notation. Using this simplification, the bound of Theorem 14 reads

$$\forall t \in [p,2]: \quad P(l_{\hat{f}} - l_{f^*}) = O\left(\left(t^* D\right)^{\frac{2}{1+\alpha}} M^{1+\frac{2}{1+\alpha}\left(\frac{1}{t^*}-1\right)} n^{-\frac{\alpha}{1+\alpha}}\right) \tag{28}$$

$\left(\text{and } P(l_{\hat{f}} - l_{f^*}) = O\left(\left(D\log M\right)^{\frac{2}{1+\alpha}} M^{\frac{\alpha-1}{\alpha+1}}\right) \text{ for } p = 1\right)$. On the other hand, the global Rademacher complexity directly leads to a bound on the supremum of the centered empirical process indexed by $\mathcal{F}$ and thus also provides a bound on the excess risk (see, e.g., Bousquet et al., 2004). Therefore, using Corollary 4, wherein we upper bound the trace of each J_m by the constant B (and subsume it under the O-notation), we have a second bound on the excess risk of the form

$$\forall t \in [p,2]: \quad P(l_{\hat{f}} - l_{f^*}) = O\left(t^* D M^{\frac{1}{t^*}} n^{-\frac{1}{2}}\right). \tag{29}$$

First consider the case where $p \geq (\log M)^*$, that is, the best choice in (28) and (29) is $t = p$. Clearly, if we hold all other parameters fixed and let n grow to infinity, the rate obtained through the local Rademacher analysis is better since $\alpha > 1$. However, it is also of interest to consider what happens when the number of kernels M and the ℓ_p ball radius D can grow with n. In general, we have a bound

on the excess risk given by the minimum of (28) and (29); a straightforward calculation shows that the local Rademacher analysis improves over the global one whenever

$$\frac{M^{\frac{1}{p}}}{D} = O(\sqrt{n}).$$

Interestingly, we note that this "phase transition" does not depend on α (i.e., the "complexity" of the individual kernels), but only on p.

If $p \leq (\log M)^*$, the best choice in (28) and (29) is $t = (\log M)^*$. In this case taking the minimum of the two bounds reads

$$\forall p \leq (\log M)^* : \quad P(l_{\hat{f}} - l_{f^*}) \leq O\Big(\min(D(\log M)n^{-\frac{1}{2}}, \big(D\log M\big)^{\frac{2}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} n^{-\frac{\alpha}{1+\alpha}}) \Big), \tag{30}$$

and the phase transition when the local Rademacher bound improves over the global one occurs for

$$\frac{M}{D\log M} = O(\sqrt{n}).$$

Finally, it is also interesting to observe the behavior of (28) and (29) as $\alpha \to \infty$. In this case, it means that only one eigenvalue is nonzero for each kernel, that is, each kernel space is one-dimensional. In other words, in this case we are in the case of "classical" aggregation of M basis functions, and the minimum of the two bounds reads

$$\forall t \in [p,2] : \quad P(l_{\hat{f}} - l_{f^*}) \leq O\Big(\min(Mn^{-1}, t^* D M^{\frac{1}{t^*}} n^{-\frac{1}{2}} \Big). \tag{31}$$

In this configuration, observe that the local Rademacher bound is $O(M/n)$ and does not depend on D, nor p, any longer; in fact, it is the same bound that one would obtain for the empirical risk minimization over the space of all linear combinations of the M base functions, without any restriction on the norm of the coefficients—the ℓ_p-norm constraint becomes void. The global Rademacher bound on the other hand, still depends crucially on the ℓ_p norm constraint. This situation is to be compared to the sharp analysis of the optimal convergence rate of convex aggregation of M functions obtained by Tsybakov (2003) in the framework of squared error loss regression, which are shown to be

$$O\left(\min\left(\frac{M}{n}, \sqrt{\frac{1}{n}\log\left(\frac{M}{\sqrt{n}}\right)} \right) \right).$$

This corresponds to the setting studied here with $D = 1, p = 1$ and $\alpha \to \infty$, and we see that the bound (30) recovers (up to log factors) in this case this sharp bound and the related phase transition phenomenon.

6.2 Discussion of Related Work

We recently learned about independent, closely related work by Suzuki (2011), which has been developed in parallel to ours. The setup considered there somewhat differs from ours: first of all, it is required that the Bayes hypothesis is contained in the class $\boldsymbol{w}^* \in H$ (which is not required in the present work); second, the conditional distribution is assumed to be expressible in terms of the Bayes hypothesis. Similar assumptions are also required in Bach (2008) in the context of sparse recovery. Finally, the analysis there is carried out for the squared loss only, while ours holds more

generally for, for example, strongly convex Lipschitz losses. However, a similarity to our setup is that an algebraic decay of the eigenvalues of the kernel matrices is assumed for the computation of the excess risk bounds and that a so-called incoherence assumption is imposed on the kernels, which is similar to our Assumption (A). Also, we do not spell out the whole analysis for inhomogeneous eigenvalue decays as Suzuki (2011) does—nevertheless, our analysis can be easily adapted to this case at the expense of longer, less-readable bounds.

We now compare the excess risk bounds of Suzuki (2011) for the case of homogeneous eigenvalue decays, that is,

$$P(l_{\hat{f}} - l_{f^*}) = O\Big((D)^{\frac{2}{1+\alpha}} M^{1+\frac{2}{1+\alpha}\left(\frac{1}{p^*}-1\right)} n^{-\frac{\alpha}{1+\alpha}} \Big),$$

to the ones shown in this paper, that is, (28)—we thereby disregard constants and the $O(n^{-1})$ terms. Roughly speaking, the proof idea in Suzuki (2011) is to exploit existing bounds on the LRC of single-kernel learning (Steinwart and Christmann, 2008) by combining Talagrand's inequality (Talagrand, 1995) and the peeling technique (van de Geer, 2000). This way the Khintchine-Kahane, which introduces a factor of $(p^*)^{\frac{2}{1+\alpha}}$ into our bounds, is avoided.

We observe that, importantly, both bounds have the same dependency in D, M, and n, although being derived by a completely different technique. Regarding the dependency in p, we observe that our bound involves a factor of $(t^*)^{\frac{2}{1+\alpha}}$ (for some $t \in [p,2]$ that is not present in the bound of Suzuki (2011). However, it can be easily shown that this factor is never of higher order than $\log(M)$ and thus can be neglected:

1. If $p \leq (\log(M))^*$, then $t = \log(M)$ is optimal in our bound so that the term $(t^*)^{\frac{2}{1+\alpha}}$ becomes $(\log(M))^{\frac{2}{1+\alpha}}$.

2. If $p \geq (\log(M))^*$, then $p^* \leq \log(M)$ so that the term $(t^*)^{\frac{2}{1+\alpha}}$ is smaller equal than $(\log(M))^{\frac{2}{1+\alpha}}$.

We can thus conclude that, besides a logarithmic factor in M as well as constants and $O(n^{-1})$ terms, our bound coincides with the rate shown in Suzuki (2011).

6.3 Discussion of Assumption (A)

Assumption **(A)** is arguably quite a strong hypothesis for the validity of our results (needed for $1 \leq p \leq 2$), which was not required for the global Rademacher bound. A similar assumption is also made in the recent works of Suzuki (2011) and Koltchinskii and Yuan (2010). In the latter paper, a related MKL algorithm using a mixture of an ℓ_1-type penalty and an empirical ℓ_2 penalty is studied (this should not be confused with $\ell_{p=1}$-norm MKL, which does not involve an empirical penalty and which, for $p = 1$, is contained in the ℓ_p-norm MKL methodology studied in this paper). Koltchinskii and Yuan (2010) derive bounds that depend on the "sparsity pattern" of the Bayes function, that is, how many coefficients w_m^* are non-zero, using an Restricted Isometry Property (RIP) assumption. If the kernel spaces are one-dimensional, in which case ℓ_1-penalized MKL reduces qualitatively to standard lasso-type methods, this assumption is known to be necessary to grant the validity of bounds taking into account the sparsity pattern of the Bayes function.[4]

4. We also mention another work by Raskutti et al. (2010), investigating the same algorithm as Koltchinskii and Yuan (2010), but employing a somewhat more restrictive assumption on the uncorrelatedness of the kernels, which corresponds to taking $c_\delta = 1$ in assumption **(A).**

In the present work, our analysis stays deliberately "agnostic" (or worst-case) with respect to the true sparsity pattern (in part because experimental evidence seems to point towards the fact that the Bayes function is not strongly sparse); correspondingly it could legitimately be hoped that the RIP condition, or Assumption **(A)**, could be substantially relaxed. Considering again the special case of one-dimensional kernel spaces and the discussion about the qualitatively equivalent case $\alpha \to \infty$ in the previous section, it can be seen that Assumption **(A)** is indeed unnecessary for bound (31) to hold, and more specifically for the rate of M/n obtained through local Rademacher analysis in this case. However, as we discussed, what happens in this specific case is that the local Rademacher analysis becomes oblivious to the ℓ_p-norm constraint, and we are left with the standard parametric convergence rate in dimension M. In other words, with one-dimensional kernel spaces, the two constraints (on the $L^2(P)$-norm of the function and on the ℓ_p block-norm of the coefficients) appearing in the definition of local Rademacher complexity are essentially not active simultaneously. Unfortunately, it is clear that this property is not true anymore for kernels of higher complexity (i.e., with a non-trivial decay rate of the eigenvalues). This is a specificity of the kernel setting as compared to combinations of a dictionary of M simple functions, and Assumption **(A)** was in effect used to "align" the two constraints. To sum up, Assumption **(A)** is used here for a different purpose from that of the RIP in sparsity analyses of ℓ_1 regularization methods; it is not clear to us at this point if this assumption is necessary or if uncorrelated variables $\boldsymbol{x}^{(m)}$ constitutes a "worst case" for our analysis. We did not suceed so far in relinquishing this assumption for $p \leq 2$, and this question remains open.

Besides the work of Suzuki (2011), there is, up to our knowledge, no previous existing analysis of the ℓ_p-MKL setting for $p > 1$; the recent works of Raskutti et al. (2010) and Koltchinskii and Yuan (2010) focus on the case $p = 1$ and on the sparsity pattern of the Bayes function. A refined analysis of ℓ_p-regularized methods in the case of combination of M basis functions was laid out by Koltchinskii (2009), also taking into account the possible soft sparsity pattern of the Bayes function. Extending the ideas underlying the latter analysis into the kernel setting is likely to open interesting developments.

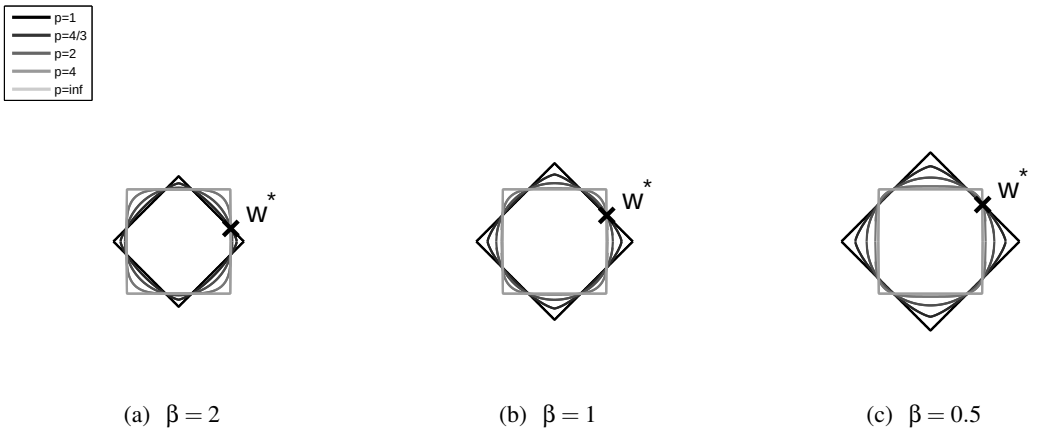
6.4 Analysis of the Impact of the Norm Parameter p on the Accuracy of ℓ_p-norm MKL

As outlined in the introduction, there is empirical evidence that the performance of ℓ_p-norm MKL crucially depends on the choice of the norm parameter p (cf. Figure 1 in the introduction). The aim of this section is to relate the theoretical analysis presented here to this empirically observed phenomenon. We believe that this phenomenon can be (at least partly) explained on base of our excess risk bound obtained in the last section. To this end we will analyze the dependency of the excess risk bounds on the chosen norm parameter p. We will show that the optimal p depends on the geometrical properties of the learning problem and that in general—depending on the true geometry—any p can be optimal. Since our excess risk bound is only formulated for $p \leq 2$, we will limit the analysis to the range $p \in [1,2]$.

To start with, first note that the choice of p only affects the excess risk bound in the factor (cf. Theorem 14 and Equation (28))

$$\nu_t := \min_{t \in [p,2]} \left(D_p t^*\right)^{\frac{2}{1+\alpha}} M^{1+\frac{2}{1+\alpha}\left(\frac{1}{t^*}-1\right)}.$$

So we write the excess risk as $P(l_{\hat{f}} - l_{f^*}) = O(\nu_t)$ and hide all variables and constants in the O-notation for the whole section (in particular the sample size n is considered a constant for the pur-



(a) $\beta = 2$ (b) $\beta = 1$ (c) $\beta = 0.5$

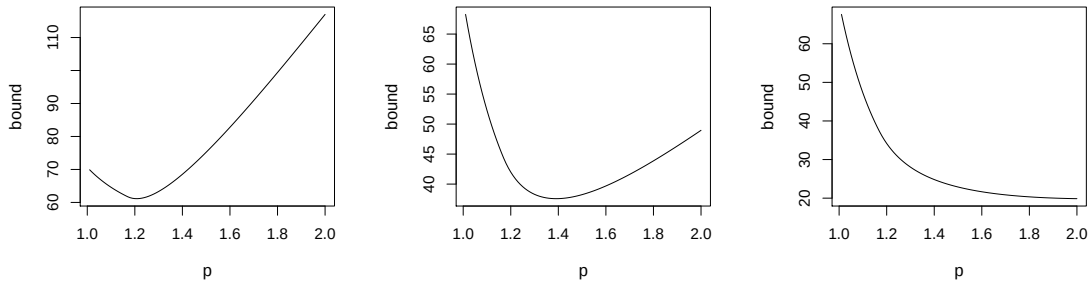
Figure 2: 2D-Illustration of the three learning scenarios analyzed in this section: LEFT: A soft sparse $\boldsymbol{w}^*$; CENTER: an intermediate non-sparse $\boldsymbol{w}^*$; RIGHT: an almost-uniformly non-sparse $\boldsymbol{w}^*$. Each scenario has a Bayes hypothesis $\boldsymbol{w}^*$ with a different soft sparsity (parametrized by β). The colored lines show the smallest ℓ_p-ball containing the Bayes hypothesis. We observe that the radii of the hypothesis classes depend on the sparsity of $\boldsymbol{w}^*$ and the parameter p.

poses of the present discussion). It might surprise the reader that we consider the term in D in the bound although it seems from the bound that it does not depend on p. This stems from a subtle reason that we have ignored in this analysis so far: D is related to the approximation properties of the class, that is, its ability to attain the Bayes hypothesis. For a "fair" analysis we should take the approximation properties of the class into account.

To illustrate this, let us assume that the Bayes hypothesis belongs to the space $\mathcal{H}$ and can be represented by $\boldsymbol{w}^*$; assume further that the block components satisfy $\|\boldsymbol{w}_m^*\|_2 = m^{-\beta}$, $m = 1,\ldots,M$, where $\beta \geq 0$ is a parameter parameterizing the "soft sparsity" of the components. For example, the cases $\beta \in \{0.5, 1, 2\}$ are shown in Figure 2 for $M = 2$ and assuming that each kernel has rank 1 (thus being isomorphic to $\mathbb{R}$). If n is large, the best bias-complexity tradeoff for a fixed p will correspond to a vanishing bias, so that the best choice of D will be close to the minimal value such that $\boldsymbol{w}^* \in H_{p,D}$, that is, $D_p = ||\boldsymbol{w}^*||_p$. Plugging in this value for D_p, the bound factor ν_p becomes

$$\nu_p := \|\boldsymbol{w}^*\|_p^{\frac{2}{1+\alpha}} \min_{t \in [p,2]} t^{* \frac{2}{1+\alpha}} M^{1+\frac{2}{1+\alpha}\left(\frac{1}{t^*}-1\right)} . \tag{32}$$

We can now plot the value ν_p as a function of p for special choices of α, M, and β. We realized this simulation for $\alpha = 2$, $M = 1000$, and $\beta \in \{0.5, 1, 2\}$, which means we generated three learning scenarios with different levels of soft sparsity parametrized by β. The results are shown in Figure 3. Note that the soft sparsity of $\boldsymbol{w}^*$ is increased from the left hand to the right hand side. We observe that in the "soft sparsest" scenario ($\beta = 2$, shown on the left-hand side) the minimum is attained for a quite small $p = 1.2$, while for the intermediate case ($\beta = 1$, shown at the center) $p = 1.4$ is optimal, and finally in the uniformly non-sparse scenario ($\beta = 2$, shown on the right-hand side) the choice of $p = 2$ is optimal (although even a higher p could be optimal, but our bound is only valid for $p \in [1,2]$).



(a) $\beta = 2$ (b) $\beta = 1$ (c) $\beta = 0.5$

Figure 3: Results of the simulation for the three analyzed learning scenarios (which were illustrated in Figure 2). The value of the bound factor ν_t is plotted as a function of p. The minimum is attained depending on the true soft sparsity of the Bayes hypothesis $\boldsymbol{w}^*$ (parametrized by β).

This means that if the true Bayes hypothesis has an intermediately dense representation, our bound gives the strongest generalization guarantees to ℓ_p-norm MKL using an intermediate choice of p. This is also intuitive: if the truth exhibits some soft sparsity but is not strongly sparse, we expect non-sparse MKL to perform better than strongly sparse MKL or the unweighted-sum kernel SVM.

6.5 An Experiment on Synthetic Data

We now present a toy experiment that is meant to check the validity of the theory presented in the previous sections. To this end, we construct learning scenarios where we know the underlying ground truth (more precisely, the ℓ_p-norm of the Bayes hypothesis) and check whether the parameter p that minimizes our bound coincides with the optimal p observed empirically, that is, when applying ℓ_p-norm MKL to the training data. Our analysis is based on the proven synthetic data described in Kloft et al. (2011) and being available from `http://mldata.org/repository/data/viewslug/mkl-toy/`. For completeness, we summarize the experimental description and the empirical results here. Note that we have extended the analysis to the whole range $p \in [1,\infty]$ (only $p \in [1,2]$ was studied in Kloft et al., 2011).

6.5.1 EXPERIMENTAL SETUP AND EMPIRICAL RESULTS

We construct six artificial data sets as described in Kloft et al. (2011), in which we vary the degree of sparsity of the true Bayes hypothesis $\boldsymbol{w}$. For each data set, we generate an $n = 50$-element, balanced sample $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from two $d = 50$-dimensional isotropic Gaussian distributions with equal covariance matrices $C = I_{d\times d}$ and equal, but opposite, means $\mu_+ = \frac{\rho}{\|\boldsymbol{w}\|_2}\boldsymbol{w}$ and $\mu_- = -\mu_+$. Figure 4 shows bar plots of the $\boldsymbol{w}$ of the various scenarios considered. The components w_i are binary valued; hence, the fraction of zero components, which we define by $\mathit{sparsity}(\boldsymbol{w}) := 1 - \frac{1}{d}\sum_{i=1}^d w_i$, is a measure for the feature sparsity of the learning problem.

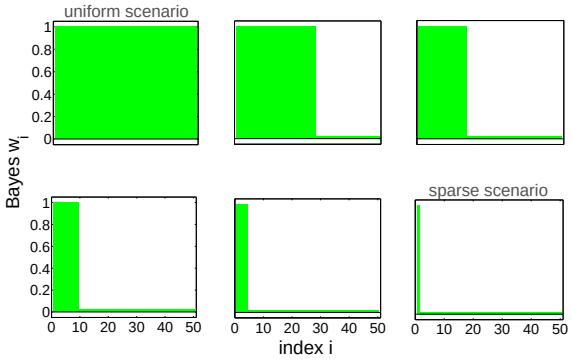


Figure 4: Toy experiment: illustration of the experimental design. We study six scenarios differing the sparsity of the Bayes hypothesis considered.

For each of the w we generate $m = 250$ data sets $\mathcal{D}_1, \ldots, \mathcal{D}_m$ fixing $\rho = 1.75$. Then, each feature is input into a linear kernel and the resulting kernel matrices are multiplicatively normalized as described in Kloft et al. (2011). Next, classification models are computed by training ℓ_p-norm MKL for $p = 1, 4/3, 2, 4, \infty$ on each $\mathcal{D}_i$. Soft margin parameters C are tuned on independent $1,000$-elemental validation sets by grid search over $C \in \left\{ 10^i \,\middle|\, i = -4, -3.5, \ldots, 0 \right\}$ (optimal Cs are attained in the interior of the grid). The relative duality gaps were optimized up to a precision of 10^{-3}. The simulation is realized for $n = 50$. We report on test errors evaluated on $1,000$-elemental independent test sets.

The results in terms of test errors are shown in Figure 5 (top). As expected, ℓ_1-norm MKL performs best and reaches the Bayes error in the sparsest scenario. In contrast, the vanilla SVM using a uniform kernel combination performs best when all kernels are equally informative. The non-sparse $\ell_{4/3}$-norm MKL variants perform best in the balanced scenarios, that is, when the noise level is ranging in the interval 64%-92%. Intuitively, the non-sparse $\ell_{4/3}$-norm MKL is the most robust MKL variant, achieving test errors of less than 12% in all scenarios. Tuning the sparsity parameter p for each experiment, ℓ_p-norm MKL achieves low test error across all scenarios.

6.5.2 BOUND

We evaluate the theoretical bound factor (32) (simply setting $\alpha = 1$) for the six learning scenarios considered. To furthermore analyze whether the p that are minimizing the bound are reflected in the empirical results, we compute the test errors of the various MKL variants again, using the setup above except that we employ a local search for finding the optimal p. The results are shown in Figure 5 (bottom). We observe a striking coincidence of the optimal p as predicted by the bound and the p that worked best empirically: In the sparsest scenario (shown on the lower right-hand side), the bound predicts $p \in [1, 1.14]$ to be optimal and indeed, in the experiments, all $p \in [1, 1.15]$ performed best (and equally well) while $p = 1.19$, already has a slightly (but significantly) worse test error—in striking match with our bounds. In the second sparsest scenario, the bound predicts $p = 1.25$ and we empirically found $p = 1.26$. In the non-sparse scenarios, intermediate values of $p \in [1, 2]$ are optimal (see Figure for details)—again we can observe a good accordance of the

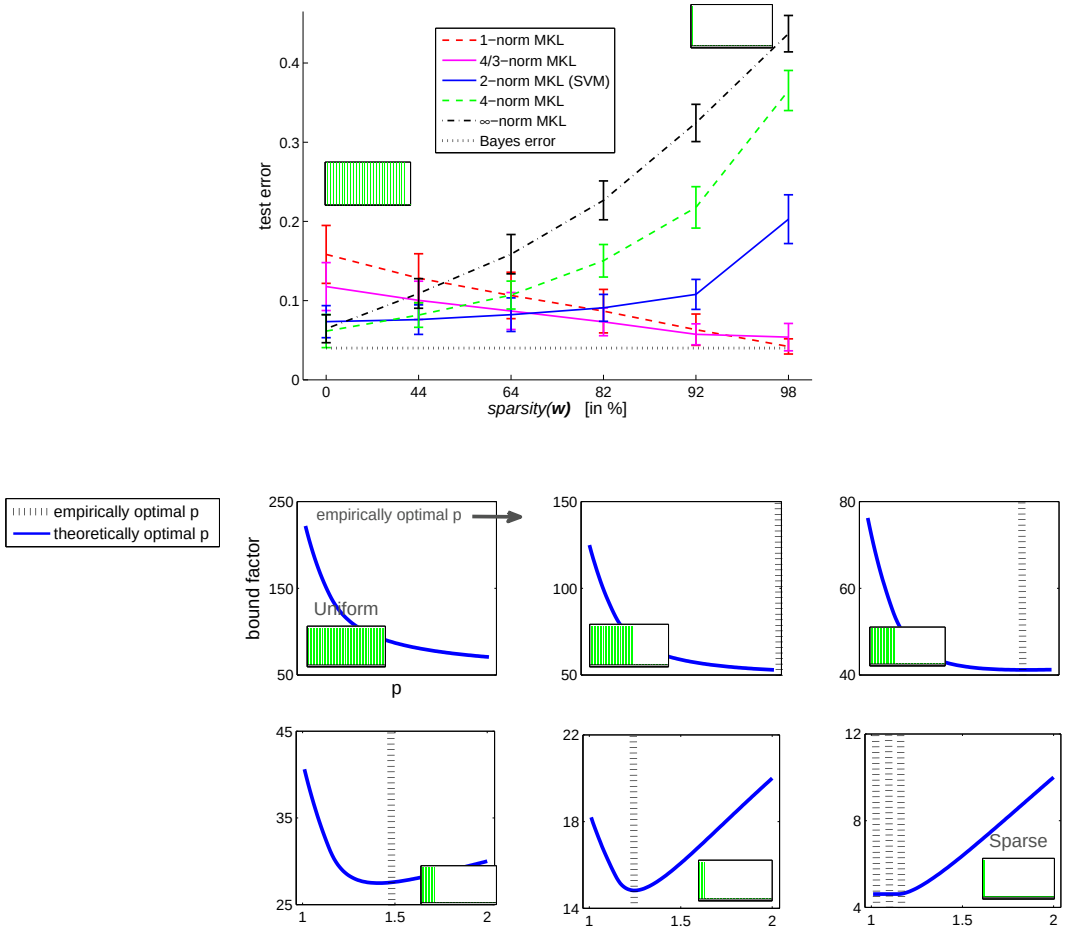


Figure 5: Toy experiment: empirical results (top) and theoretical bounds (bottom).

empirical and theoretical values. In the extreme case, that is, the uniform scenario, the bound indicates a p that lies well beyond the valid interval of the bound (i.e., $p > 2$) and this is also what we observe empirically: $p \in [4, \infty]$ worked best in our experiments.

6.5.3 SUMMARY AND DISCUSSION

We can conclude that the empirical results indicate the validity of our theory: the theoretical bounds reflect the empirically observed optimal p in the idealized setup where we know the underlying ground true, that is, the ℓ_p-norm of the Bayes hypothesis. We also observed that the optimality of a particular p strongly depends on the geometry of the learning task: the sparsity of the underlying Bayes hypothesis $\boldsymbol{w}$. This raises the question into which scenarios practical applications fall. For example, do we rather encounter a "sparse" or non-sparse scenario in bioinformatics? However, this investigation is beyond the scope of this paper (see Chapter 5 in Kloft (2011) for an analysis aiming in that direction).

The results of our analysis are especially surprising, when recalling the result of Suzuki (2011) discussed in Section 6.2. For the setup of homogeneous eigenvalue decay of the kernels as considered in the toy experiment setup here, their bound is optimal for $p = 1$, regardless of the sparsity of the Bayes hypothesis. This is counter-intuitive and in strong contrast to our empirical analysis on synthetic data, where the optimality of a certain value of the norm parameter p crucially depends on the sparsity of the Bayes hypothesis. At this point we have no explanation for this behavior and this leaves an open issue for relating theory to empirical results. The analysis carried out in this paper may serve as a starting point for subsequent analyses aiming in that direction.

7. Conclusion

We derived a sharp upper bound on the local Rademacher complexity of ℓ_p-norm multiple kernel learning under the assumption of uncorrelated kernels. We also proved a lower bound that matches the upper one and shows that our result is tight. Using the local Rademacher complexity bound, we derived an excess risk bound that attains the fast rate of $O(n^{-\frac{\alpha}{1+\alpha}})$, where α is the minimum eigenvalue decay rate of the individual kernels.

In a practical case study, we found that the optimal value of that bound depends on the true Bayes-optimal kernel weights. If the true weights exhibit soft sparsity but are not strongly sparse, then the generalization bound is minimized for an intermediate p. This is not only intuitive but also supports empirical studies showing that sparse MKL ($p = 1$) rarely works in practice, while some intermediate choice of p can improve performance.

Of course, this connection is only valid if the optimal kernel weights are likely to be non-sparse in practice. Indeed, related research points in that direction. For example, already weak connectivity in a causal graphical model may be sufficient for all variables to be required for optimal predictions, and even the prevalence of sparsity in causal flows is being questioned (e.g., for the social sciences Gelman, 2010, argues that "There are (almost) no true zeros").

Finally, we note that there seems to be a certain preference for sparse models in the scientific community. However, previous MKL research has shown that non-sparse models may improve quite impressively over sparse ones in practical applications. The present analysis supports this by showing that the reason for this might be traced back to non-sparse MKL attaining better generalization bounds in non-sparse learning scenarios. We remark that this point of view is also supported by related analyses.

For example, it was shown by Leeb and Pötscher (2008) in a fixed design setup that any sparse estimator (i.e., satisfying the oracle property of correctly predicting the zero values of the true target $\boldsymbol{w}^*$) has a maximal scaled mean squared error (MSMSE) that diverges to ∞. This is somewhat suboptimal since, for example, least-squares regression has a converging MSMSE. Although this is an asymptotic result, it might also be one of the reasons for finding excellent (non-asymptotic) results in non-sparse MKL. In another, recent study of Xu et al. (2008), it was shown that no sparse algorithm can be algorithmically stable. This is noticeable because algorithmic stability is connected with generalization error (Bousquet and Elisseeff, 2002).

Acknowledgments

We greatly thank Peter L. Bartlett, Klaus-Robert Müller, and Taiji Suzuki for their helpful comments on the manuscript.

MK was supported by the German Science Foundation (DFG MU 987/6-1, RA 1894/1-1) and by the World Class University Program through the National Research Foundation of Korea funded by the Korean Ministry of Education, Science, and Technology, under Grant R31-10008. GB was supported by the European Community under the grant agreement 247022 (MASH Project). Both authors were supported by the European Community's 7th Framework Programme under the PASCAL2 Network of Excellence (ICT-216886).

Appendix A. Relation of MKL to Block-Norm Formulation

For completeness, we show in this appendix the relation of kernel weights formulation of MKL to block-norm formulation.

A.1 The Case $p \in [1,2]$

We show that denoting $\boldsymbol{w} = (\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(M)})$, for any $q \in [1,\infty]$, the hypothesis class

$$\left\{ f : x \mapsto \sum_{m=1}^{M} \left\langle \boldsymbol{w}_m, \sqrt{\theta_m}\phi_m(x) \right\rangle \;\middle|\; \|\boldsymbol{w}\|_2 \le D,\ \|\boldsymbol{\theta}\|_q \le 1 \right\}, \tag{33}$$

is identical to the block norm class

$$H_{p,D,M} = \left\{ f : x \mapsto \langle \boldsymbol{w}, \phi(x) \rangle \;\middle|\; \|\boldsymbol{w}\|_{2,p} \le D \right\} \tag{34}$$

where $p := \frac{2q}{q+1}$. This is known since Micchelli and Pontil (2005). To this end, first we rewrite (33) as

$$H_{p,D,M} = \left\{ f : x \mapsto \langle \boldsymbol{w}, \phi(x) \rangle \;\middle|\; \sum_{m=1}^{M} \frac{\|\boldsymbol{w}_m\|_2^2}{\theta_m} \le D^2,\ \|\boldsymbol{\theta}\|_q \le 1 \right\}. \tag{35}$$

However, solving

$$\inf_{\boldsymbol{\theta}} \quad \frac{1}{2} \sum_{m=1}^{M} \frac{\|\boldsymbol{w}_m\|_2^2}{\theta_m}, \quad \text{s.t.} \quad \|\boldsymbol{\theta}\|_q \le 1$$

for fixed $\boldsymbol{w} > \boldsymbol{0}$, the optimal $\boldsymbol{\theta}$ is attained at

$$\theta_m = \frac{\|\boldsymbol{w}_m\|_2^{\frac{2}{q+1}}}{\left(\sum_{m'=1}^{M} \|\boldsymbol{w}_{m'}\|_2^{\frac{2q}{q+1}} \right)^{1/q}}, \qquad \forall m = 1, \ldots, M.$$

Plugging the latter into (35), we obtain (34) with $p = \frac{2q}{q+1}$, which was to show.

A.2 The Case $p \in]2, \infty]$

Even if $p > 2$, we can obtain an alternative formulation of the block norm MKL problem, as shown in Aflalo et al. (2011), by the definition of the dual norm $\|\cdot\|_*$ of a norm $\|\cdot\|$, that is, $\|x\|_* = \sup_{\boldsymbol{y}} \langle \boldsymbol{x}, \boldsymbol{y} \rangle - \|y\|$, it holds

$$\|\boldsymbol{w}\|_{2,p}^2 \;=\; \left\| \left(\|\boldsymbol{w}_m\|_2^2 \right)_{m=1}^M \right\|_{p/2} = \sup_{\boldsymbol{\theta}: \|\boldsymbol{\theta}\|_{(p/2)^*} \leq 1} \; \sum_{m=1}^M \theta_m \|\boldsymbol{w}_m\|_2^2 \,.$$

Thus defining $q := (p/2)^*$ we can obtain a learning-the-kernel MKL formulation from the above equation. A difference to the case $p < 2$ lies in the kernel weights θ_m appearing in the nominator instead of the denominator.

Appendix B. Lemmata and Proofs

The following result gives a block-structured version of Hölder's inequality (e.g., Steele, 2004).

Lemma 15 (Block-structured Hölder inequality)**.** *Let* $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}),\ \boldsymbol{y} = (\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(m)}) \in \mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_M$*. Then, for any* $p \geq 1$*, it holds*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \|\boldsymbol{x}\|_{2,p} \|\boldsymbol{y}\|_{2,p^*} \,.$$

Proof By the Cauchy-Schwarz inequality (C.-S.), we have for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$:

$$\begin{aligned} \langle \boldsymbol{x}, \boldsymbol{y} \rangle \quad &= \quad \sum_{m=1}^M \langle \boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)} \rangle \overset{\text{C.-S.}}{\leq} \sum_{m=1}^M \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \\ &= \quad \left\langle (\|\boldsymbol{x}^{(1)}\|_2, \ldots, \|\boldsymbol{x}^{(M)}\|_2), (\|\boldsymbol{y}^{(1)}\|_2, \ldots, \|\boldsymbol{y}^{(M)}\|_2) \right\rangle. \\ &\overset{\text{Hölder}}{\leq} \quad \|\boldsymbol{x}\|_{2,p} \|\boldsymbol{y}\|_{2,p^*} \end{aligned}$$

■

Proof of Lemma 3 (Rosenthal + Young) It is clear that the result trivially holds for $\frac{1}{2} \leq p \leq 1$ with $C_q = 1$ by Jensen's inequality . In the case $p \geq 1$, we apply Rosenthal's inequality (Rosenthal, 1970) to the sequence $X_1, \ldots, X_n$ thereby using the optimal constants computed in Ibragimov and Sharakhmetov (2001), that are, $C_q = 2$ ($q \leq 2$) and $C_q = \mathbb{E}Z^q$ ($q \geq 2$), respectively, where Z is a random variable distributed according to a Poisson law with parameter $\lambda = 1$. This yields

$$\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i \right)^q \leq C_q \max\left(\frac{1}{n^q} \sum_{i=1}^n \mathbb{E}X_i^q, \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^q \right). \tag{36}$$

By using that $X_i \leq B$ holds almost surely, we could readily obtain a bound of the form $\frac{B^q}{n^{q-1}}$ on the first term. However, this is loose and for $q = 1$ does not converge to zero when $n \to \infty$. Therefore,

we follow a different approach based on Young's inequality (e.g., Steele, 2004):

$$\begin{aligned}\frac{1}{n^q}\sum_{i=1}^n \mathbb{E}X_i^q &\leq \left(\frac{B}{n}\right)^{q-1}\frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i \\ &\overset{\text{Young}}{\leq} \frac{1}{q^*}\left(\frac{B}{n}\right)^{q^*(q-1)} + \frac{1}{q}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i\right)^q \\ &= \frac{1}{q^*}\left(\frac{B}{n}\right)^q + \frac{1}{q}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i\right)^q.\end{aligned}$$

It thus follows from (36) that for all $q \geq \frac{1}{2}$

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^q \leq C_q\left(\left(\frac{B}{n}\right)^q + \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i\right)^q\right),$$

where C_q can be taken as 2 ($q \leq 2$) and $\mathbb{E}Z^q$ ($q \geq 2$), respectively, where Z is Poisson-distributed. In the subsequent Lemma 16 we show $\mathbb{E}Z^q \leq (q+e)^q$. Clearly, for $q \geq \frac{1}{2}$ it holds $q+e \leq qe+eq = 2eq$ so that in any case $C_q \leq \max(2, 2eq) \leq 2eq$, which concludes the result. ■

We use the following Lemma gives a handle on the q-th moment of a Poisson-distributed random variable and is used in the previous Lemma.

Lemma 16. *For the q-moment of a random variable Z distributed according to a Poisson law with parameter $\lambda = 1$, the following inequality holds for all $q \geq 1$:*

$$\mathbb{E}Z^q \overset{def.}{=} \frac{1}{e}\sum_{k=0}^\infty \frac{k^q}{k!} \leq (q+e)^q.$$

Proof We start by decomposing $\mathbb{E}Z^q$ as follows:

$$\begin{aligned}\mathbb{E}^q &= \frac{1}{e}\left(0 + \sum_{k=1}^q \frac{k^q}{k!} + \sum_{k=q+1}^\infty \frac{k^q}{k!}\right) \\ &= \frac{1}{e}\left(\sum_{k=1}^q \frac{k^{q-1}}{(k-1)!} + \sum_{k=q+1}^\infty \frac{k^q}{k!}\right) \\ &\leq \frac{1}{e}\left(q^q + \sum_{k=q+1}^\infty \frac{k^q}{k!}\right) \qquad (37)\end{aligned}$$

(38)

Note that by Stirling's approximation it holds $k! = \sqrt{2\pi}e^{\tau_k}k\left(\frac{k}{e}\right)^q$ with $\frac{1}{12k+1} < \tau_k < \frac{1}{12k}$ for all q. Thus

$$\begin{aligned}
\sum_{k=q+1}^{\infty}\frac{k^q}{k!} &= \sum_{k=q+1}^{\infty}\frac{1}{\sqrt{2\pi}e^{\tau_k}k}e^k k^{-(k-q)} \\
&= \sum_{k=1}^{\infty}\frac{1}{\sqrt{2\pi}e^{\tau_{k+q}}(k+q)}e^{k+q}k^{-k} \\
&= e^q\sum_{k=1}^{\infty}\frac{1}{\sqrt{2\pi}e^{\tau_{k+q}}(k+q)}\left(\frac{e}{k}\right)^k \\
&\overset{(*)}{\leq} e^q\sum_{k=1}^{\infty}\frac{1}{\sqrt{2\pi}e^{\tau_k}k}\left(\frac{e}{k}\right)^k \\
&\overset{\text{Stirling}}{=} e^q\sum_{k=1}^{\infty}\frac{1}{k!} \\
&= e^{q+1}
\end{aligned}$$

where for $(*)$ note that $e^{\tau_k}k \leq e^{\tau_{k+q}}(k+q)$ can be shown by some algebra using $\frac{1}{12k+1} < \tau_k < \frac{1}{12k}$. Now by (37)

$$\mathbb{E}Z^q = \frac{1}{e}\left(q^q + e^{q+1}\right) \leq q^q + e^q \leq (q+e)^q,$$

which was to show. ■

Lemma 17. *For any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m_+$ it holds for all $q \geq 1$*

$$\|\boldsymbol{a}\|_q + \|\boldsymbol{b}\|_q \leq 2^{1-\frac{1}{q}}\|\boldsymbol{a}+\boldsymbol{b}\|_q \leq 2\|\boldsymbol{a}+\boldsymbol{b}\|_q.$$

Proof Let $\boldsymbol{a} = (a_1, \ldots, a_m)$ and $\boldsymbol{b} = (b_1, \ldots, b_m)$. Because all components of $\boldsymbol{a}, \boldsymbol{b}$ are nonnegative, we have

$$\forall i = 1, \ldots, m: \; a_i^q + b_i^q \leq \left(a_i + b_i\right)^q$$

and thus

$$\|\boldsymbol{a}\|_q^q + \|\boldsymbol{b}\|_q^q \leq \|\boldsymbol{a}+\boldsymbol{b}\|_q^q. \tag{39}$$

We conclude by ℓ_q-to-ℓ_1 conversion (see (20))

$$\begin{aligned}
\|\boldsymbol{a}\|_q + \|\boldsymbol{b}\|_q &= \left\|\left(\|\boldsymbol{a}\|_q, \|\boldsymbol{b}\|_q\right)\right\|_1 \overset{(20)}{\leq} 2^{1-\frac{1}{q}}\left\|\left(\|\boldsymbol{a}\|_q, \|\boldsymbol{b}\|_q\right)\right\|_q = 2^{1-\frac{1}{q}}\left(\|\boldsymbol{a}\|_q^q + \|\boldsymbol{b}\|_q^q\right)^{\frac{1}{q}} \\
&\overset{(39)}{\leq} 2^{1-\frac{1}{q}}\|\boldsymbol{a}+\boldsymbol{b}\|_q,
\end{aligned}$$

which completes the proof. ■

References

J. Aflalo, A. Ben-Tal, C. Bhattacharyya, J. S. Nath, and S. Raman. Variable sparsity kernel learning—algorithms and applications. *Journal of Machine Learning Research*, 12:565–592, Feb 2011.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9: 1179–1225, 2008.

F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proc. 21st ICML*. ACM, 2004.

P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, Nov. 2002.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. (Was Department of Statistics, U.C. Berkeley Technical Report number 638, 2003).

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Annals of Statistics*, 36(2):489–531, 2008.

R. R. Bouckaert, E. Frank, M. A. Hall, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. WEKA—experiences with a java open-source project. *Journal of Machine Learning Research*, 11:2533–2541, 2010.

O. Bousquet and A. Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, March 2002. ISSN 1532-4435.

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer Berlin / Heidelberg, 2004.

C. Cortes. Invited talk: Can learning kernels help performance? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1:1–1:1, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1.

C. Cortes, A. Gretton, G. Lanckriet, M. Mohri, and A. Rostamizadeh. Proceedings of the NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels, 2008. URL `http://www.cs.nyu.edu/learning_kernels`.

C. Cortes, M. Mohri, and A. Rostamizadeh. L2 regularization for learning kernels. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2009.

C. Cortes, M. Mohri, and A. Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings, 27th ICML*, 2010.

P. V. Gehler and S. Nowozin. Let the kernel figure it out: Principled learning of pre-processing for kernel classifiers. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 06 2009.

A. Gelman. Causality and statistical learning. *American Journal of Sociology*, 0, 2010.

R. Ibragimov and S. Sharakhmetov. The best constant in the rosenthal inequality for nonnegative random variables. *Statistics & Probability Letters*, 55(4):367 – 376, 2001. ISSN 0167-7152.

J.-P. Kahane. *Some Random Series of Functions*. Cambridge University Press, 2nd edition, 1985.

M. Kloft. *ℓ_p-Norm Multiple Kernel Learning*. PhD thesis, Berlin Institute of Technology, Oct 2011. URL `http://opus.kobv.de/tuberlin/volltexte/2011/3239/`.

M. Kloft, U. Brefeld, P. Laskov, and S. Sonnenburg. Non-sparse multiple kernel learning. In *Proc. of the NIPS Workshop on Kernel Learning: Automatic Selection of Kernels*, dec 2008.

M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien. Efficient and accurate lp-norm multiple kernel learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 997–1005. MIT Press, 2009.

M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. ℓ_p-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, Mar 2011.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 45(1):7–57, 2009.

V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *Annals of Statistics*, 38(6): 3660–3695, 2010.

S. Kwapién and W. A. Woyczyński. *Random Series and Stochastic Integrals: Single and Multiple*. Birkhäuser, Basel and Boston, M.A., 1992.

G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett, and M. I. Jordan. Learning the kernel matrix with semi-definite programming. *JMLR*, 5:27–72, 2004.

H. Leeb and B. M. Pötscher. Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, 142:201–211, 2008.

C. McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

S. Mendelson. On the performance of kernel classes. *J. Mach. Learn. Res.*, 4:759–771, December 2003.

C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Neural Networks*, 12(2):181–201, May 2001.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *CoRR*, abs/1008.3654, 2010.

H. Rosenthal. On the subspaces of L_p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 8:273–303, 1970.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

J. R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(03):417–424, 1980. doi: 10.1017/S0140525X00005756. URL `http://dx.doi.org/10.1017/S0140525X00005756`.

S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, July 2006.

N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In G. Lugosi and H.-U. Simon, editors, *COLT*, volume 4005 of *Lecture Notes in Computer Science*, pages 169–183. Springer, 2006.

J. M. Steele. *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge University Press, New York, NY, USA, 2004. ISBN 052154677X.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

M. Stone. Cross-validatory choice and assessment of statistical predictors (with discussion). *Journal of the Royal Statistical Society*, B36:111–147, 1974.

T. Suzuki. Unifying framework for fast learning rate of non-sparse multiple kernel learning. In J. Shawe-Taylor, R. Zemel, P. L. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*. 2011. To appear.

M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathmatiques de L'IHS*, 81:73–205, 1995.

A. Tsybakov. Optimal rates of aggregation. In B. Schölkopf and M. Warmuth, editors, *Computational Learning Theory and Kernel Machines (COLT-2003)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, 2003.

S. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47(6):2516–2532, 2001.

H. Xu, S. Mannor, and C. Caramanis. Sparse algorithms are not stable: A no-free-lunch theorem. In *Proceedings of the 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 1299 –1303, 2008.

Y. Ying and C. Campbell. Generalization bounds for learning the kernel problem. In *COLT*, 2009.