The Local Rademacher Complexity of ℓ_p-Norm Multiple Kernel Learning

Marius Kloft*
Machine Learning Laboratory
TU Berlin, Germany
kloft@tu-berlin.de

Gilles Blanchard
Department of Mathematics
University of Potsdam, Germany
gilles.blanchard@math.uni-potsdam.de

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.

1 Introduction

Kernel methods [24, 21] allow to obtain nonlinear learning machines from simpler, linear ones; nowadays they can almost completely be applied out-of-the-box [3]. Nevertheless, after more than a decade of research it still remains an unsolved problem to find the best abstraction or *kernel* for a problem at hand. Most frequently, the kernel is selected from a candidate set according to its generalization performance on a validation set. Clearly, the performance of such an algorithm is limited by the best kernel in the set. Unfortunately, in the current state of research, there is little hope that in the near future a *machine* will be able to automatically find—or even engineer—the best kernel for a particular problem at hand [25]. However, by restricting to a less general problem, can we hope to achieve the automatic kernel selection?

In the seminal work of Lanckriet et al. [18] it was shown that learning a support vector machine (SVM) [9] *and* a convex kernel combination at the same time is computationally feasible. This approach was entitled *multiple kernel learning* (MKL). Research in the subsequent years focused on speeding up the initially demanding optimization algorithms [22, 26]—ignoring the fact that empirical evidence for the superiority of MKL over trivial baseline approaches (not optimizing the kernel) was missing. In 2008, negative results concerning the accuracy of MKL in practical applications accumulated: at the NIPS 2008 MKL workshop [6] several researchers presented empirical evidence showing that traditional MKL rarely helps in practice and frequently is outperformed by a regular SVM using a uniform kernel combination, see http://videolectures.net/lkasok08_whistler/. Subsequent research (e.g., [10]) revealed further negative evidence and peaked in the provocative question "Can learning kernels help performance?" posed by Corinna Cortes in an invited talk at ICML 2009 [5].

Consequently, despite all the substantial progress in the field of MKL, there remained an unsatisfied need for an approach that is really useful for practical applications: a model that has a good chance of improving the accuracy (over a plain sum kernel). A first step towards a model of kernel learning

*Marius Kloft is also with Friedrich Miescher Laboratory, Max Planck Society, Tübingen. A part of this work was done while Marius Kloft was with UC Berkeley, USA, and Gilles Blanchard was with Weierstraß Institute for Applied Analysis and Stochastics, Berlin.

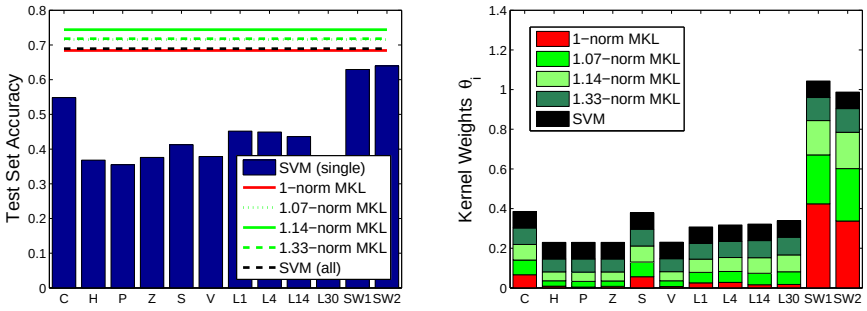


Figure 1: Result of a typical ℓ_p-norm MKL experiment in terms of accuracy (LEFT) and kernel weights output by MKL (RIGHT).

that is useful for practical applications was made in [7, 13, 14]: by imposing an ℓ_q-norm penalty ($q > 1$) rather than an ℓ_1-norm one on the kernel combination coefficients. This 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 $\left\{k = \sum_{m=1}^{M} \theta_m k_m \;\middle|\; \|\boldsymbol{\theta}\|_q \leq 1,\ \boldsymbol{\theta} \geq 0\right\}$. A conceptual milestone going back to the work of [1] and [20] is that this multi-kernel class can equivalently be represented as a block-norm regularized linear class in the product RKHS:

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

where there is a one-to-one mapping of $q \in [1, \infty]$ to $p \in [1, 2]$ given by $p = \frac{2q}{q+1}$.

In Figure 1, we show exemplary results of an ℓ_p-norm MKL experiment, achieved on the protein fold prediction dataset used in [4] (see supplementary material A for experimental details). We first observe that, as expected, ℓ_p-norm MKL enforces strong sparsity in the coefficients θ_m when $p = 1$ and no sparsity at all otherwise (but various degrees of soft sparsity for intermediate p). Crucially, the performance (as measured by the test error) is not monotonic as a function of p; $p = 1$ (sparse MKL) yields the same performance as the regular SVM using a uniform kernel combination, but optimal performance is attained for some intermediate value of p—namely, $p = 1.14$. This is a strong empirical motivation to study theoretically the performance of ℓ_p-MKL beyond the limiting cases $p = 1$ and $p = \infty$.

Clearly, the complexity of (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 hypothesis class (1). An analysis of this model, based on global Rademacher complexities, was developed by [8] for special cases of p. 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 that cover the whole range of $p \in [1, \infty]$.

Outline of the contributions. This paper makes the following contributions:

- An upper bound on the local Rademacher complexity of ℓ_p-norm MKL is 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 (previous bounds for ℓ_p-norm MKL only achieved $O(M^{\frac{1}{p^*}} n^{-\frac{1}{2}})$.
- A lower bound is shown that beside 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.

Furthermore, we also present a simpler, more general proof of the global Rademacher bound shown in [8] (at the expense of a slightly worse constant). A comparison of the rates obtained with local and global Rademacher analysis is carried out in Section 3.

Notation. We abbreviate $H_p = H_{p,D} = H_{p,D,M}$ if clear from the context. We denote the (normalized) kernel matrices corresponding to k and k_m by K and K_m, respectively, i.e., the ijth entry of K is $\frac{1}{n}k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Also, we denote $\boldsymbol{u} = (\boldsymbol{u}^{(m)})_{m=1}^M = (\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)}) \in \mathcal{H} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_M$. Furthermore, let P be a probability measure on $\mathcal{X}$ i.i.d. generating the data $x_1, \ldots, x_n$ and denote by $\mathbb{E}$ the corresponding expectation operator. We work with operators in Hilbert spaces and will use instead of the usual vector/matrix notation $\phi(x)\phi(x)^\top$ the tensor notation $\phi(x) \otimes \phi(x) \in \mathrm{HS}(\mathcal{H})$, which is a Hilbert-Schmidt operator $\mathcal{H} \mapsto \mathcal{H}$ defined as $(\phi(x) \otimes \phi(x))\boldsymbol{u} = \langle \phi(x), \boldsymbol{u} \rangle \, \phi(x)$. The space $\mathrm{HS}(\mathcal{H})$ of Hilbert-Schmidt operators on $\mathcal{H}$ is itself a Hilbert space, and the expectation $\mathbb{E}\phi(x) \otimes \phi(x)$ is well-defined and belongs to $\mathrm{HS}(\mathcal{H})$ as soon as $\mathbb{E}\left\|\phi(x)\right\|^2$ is finite, which will always be assumed. We denote by $J = \mathbb{E}\phi(x) \otimes \phi(x)$ and $J_m = \mathbb{E}\phi_m(x) \otimes \phi_m(x)$ the uncentered covariance operators corresponding to variables $\phi(x)$ and $\phi_m(x)$, respectively; it holds that $\mathrm{tr}(J) = \mathbb{E}\left\|\phi(x)\right\|_2^2$ and $\mathrm{tr}(J_m) = \mathbb{E}\left\|\phi_m(x)\right\|_2^2$.

Global Rademacher Complexities We first review global Rademacher complexities (GRC) in multiple kernel learning. Let $x_1, \ldots, x_n$ be an i.i.d. sample drawn from P. The global Rademacher complexity is defined as $R(H_p) = \mathbb{E}\sup_{f_{\boldsymbol{w}} \in H_p} \langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^n \sigma_i \phi(x_i) \rangle$, where $(\sigma_i)_{1 \le i \le n}$ is an i.i.d. family (independent of $\phi(x_i)$) of Rademacher variables (random signs). Its empirical counterpart is denoted by $\widehat{R}(H_p) = \mathbb{E}\big[R(H_p)\big|x_1, \ldots, x_n\big] = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f_{\boldsymbol{w}} \in H_p} \langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^n \sigma_i \phi(x_i) \rangle$.

In the recent paper of [8] it was shown $\widehat{R}(H_p) \le \frac{D}{n}\left(cp^* \left\| \left(\mathrm{tr}(K_m)\right)_{m=1}^M \right\|_{\frac{p^*}{2}}\right)^{1/2}$ for $p \in [1,2]$ and p^* being an integer (where $c = 23/44$ and $p^* := \frac{p}{p-1}$ is the conjugated exponent). This bound is tight and improves a series of loose results that were given for $p = 1$ in the past (see [8] and references therein). In fact, the above result can be extended to the whole range of $p \in [1, \infty]$ (in the supplementary material we present a quite simple proof using $c = 1$):

Proposition 1 (GLOBAL RADEMACHER COMPLEXITY BOUND)**.** *For any $p \ge 1$ the empirical version of global Rademacher complexity of the multi-kernel class H_p can be bounded as*

$$\widehat{R}(H_p) \le \min_{t \in [p,\infty]} D\sqrt{\frac{t^*}{n}\left\| \left(\frac{1}{n}\mathrm{tr}(K_m)\right)_{m=1}^M \right\|_{\frac{t^*}{2}}}.$$

Interestingly, the above GRC bound is not monotonic in p and thus the minimum is not always attained for $t := p$.

2 The Local Rademacher Complexity of Multiple Kernel Learning

Let $x_1, \ldots, x_n$ be an i.i.d. sample drawn from P. We define the *local* Rademacher complexity (LRC) of H_p as $R_r(H_p) = \mathbb{E}\sup_{f_{\boldsymbol{w}} \in H_p : Pf_{\boldsymbol{w}}^2 \le r} \langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^n \sigma_i \phi(x_i) \rangle$, where $Pf_{\boldsymbol{w}}^2 := \mathbb{E}(f_{\boldsymbol{w}}(\phi(x)))^2$. Note that it subsumes the global RC as a special case for $r = \infty$. We will also use the following assumption in the bounds for the case $p \in [1,2]$:

Assumption (U) (no-correlation). *Let $x \sim P$. The Hilbert space valued random variables $\phi_1(x), \ldots, \phi_M(x)$ are said to be (pairwise) uncorrelated if for any $m \ne m'$ and $\boldsymbol{w} \in \mathcal{H}_m$, $\boldsymbol{w}' \in \mathcal{H}_{m'}$, the real variables $\langle \boldsymbol{w}, \phi_m(x) \rangle$ and $\langle \boldsymbol{w}', \phi_{m'}(x) \rangle$ are uncorrelated.*

For example, if $\mathcal{X} = \mathbb{R}^M$, the above means that the 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 also considered by [23] (for sparse MKL). To state the bounds, note that covariance operators enjoy discrete eigenvalue-eigenvector decompositions $J = \mathbb{E}\phi(x) \otimes \phi(x) = \sum_{j=1}^\infty \lambda_j \boldsymbol{u}_j \otimes \boldsymbol{u}_j$ 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 $(\boldsymbol{u}_j)_{j \ge 1}$ and $(\boldsymbol{u}_j^{(m)})_{j \ge 1}$ form orthonormal bases of $\mathcal{H}$ and $\mathcal{H}_m$, respectively. We are now equipped to state our main results:

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

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

Theorem 3 (LOCAL RADEMACHER COMPLEXITY BOUND, $p \in [2,\infty]$). *For any $p \in [2,\infty]$,*

$$R_r(H_p) \;\leq\; \min_{t\in[p,\infty]} \sqrt{\frac{2}{n}\sum_{j=1}^{\infty} \min(r, D^2 M^{\frac{2}{t^*}-1}\lambda_j)}.$$

It is interesting to compare the above bounds for the special case $p=2$ with the ones of Bartlett et al. [2]. The main term of the bound of Theorem 3 (taking $t=p=2$) is then essentially determined by $O\big(\big(\frac{1}{n}\sum_{m=1}^{M}\sum_{j=1}^{\infty}\min(r,\lambda_j^{(m)})\big)^{1/2}\big)$. If the variables $(\phi_m(x))$ are centered and uncorrelated, this is equivalently of order $O\big(\big(\frac{1}{n}\sum_{j=1}^{\infty}\min(r,\lambda_j)\big)^{1/2}\big)$ because $\operatorname{spec}(J) = \bigcup_{m=1}^{M}\operatorname{spec}(J_m)$; that is, $\{\lambda_i, i\geq 1\} = \bigcup_{m=1}^{M}\{\lambda_i^{(m)}, i\geq 1\}$; this rate is also what we would obtain through Theorem 3, so both bounds on the LRC recover the rate shown in [2] for the special case $p=2$.

It is also interesting to study the case $p=1$: by using $t=(\log(M))^*$ in Theorem 2, we obtain the bound $R_r(H_1) \leq \big(\frac{16}{n}\big\|\big(\sum_{j=1}^{\infty}\min\big(rM, e^3D^2(\log M)^2\lambda_j^{(m)}\big)\big)_{m=1}^{M}\big\|_{\infty}\big)^{1/2} + \frac{\sqrt{B}e^{\frac{3}{2}}D\log(M)}{n}$, for all $M \geq e^2$. We now turn to proving Theorem 2. the proof of Theorem 3 is straightforward and shown in the supplementary material C.

Proof of Theorem 2. . Note that it suffices to prove the result for $t=p$ as trivially $\|\boldsymbol{w}\|_{2,t} \leq \|\boldsymbol{w}\|_{2,p}$ holds for all $t\geq p$ so that $H_p \subseteq H_t$ and therefore $R_r(H_p) \leq R_r(H_t)$.

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 $\widetilde{H_p} = \big\{\widetilde{f}_{\boldsymbol{w}} \;\big|\; \|\boldsymbol{w}\|_{2,p}\leq D\big\}$, wherein $\widetilde{f}_{\boldsymbol{w}} : x \mapsto \langle \boldsymbol{w}, \widetilde{\phi}(x)\rangle$, and $\widetilde{\phi}(x) := \phi(x) - \mathbb{E}\phi(x)$. We start the proof by noting that $\widetilde{f}_{\boldsymbol{w}}(x) = f_{\boldsymbol{w}}(x) - \langle \boldsymbol{w}, \mathbb{E}\phi(x)\rangle = f_{\boldsymbol{w}}(x) - \mathbb{E}\,\langle \boldsymbol{w}, \phi(x)\rangle = f_{\boldsymbol{w}}(\phi(x)) - \mathbb{E}f_{\boldsymbol{w}}(\phi(x))$, so that, by the bias-variance decomposition, it holds that

$$Pf_{\boldsymbol{w}}^2 = \mathbb{E}f_{\boldsymbol{w}}(x)^2 = \mathbb{E}\,(f_{\boldsymbol{w}}(x) - \mathbb{E}f_w(x))^2 + (\mathbb{E}f_w(\phi(x)))^2 = P\widetilde{f}_{\boldsymbol{w}}^2 \;+\; \big(Pf_{\boldsymbol{w}}\big)^2. \tag{2}$$

Furthermore we note that by Jensen's inequality

$$\begin{aligned}\big\|\mathbb{E}\phi(x)\big\|_{2,p^*} \;&=\; \left(\sum_{m=1}^{M}\big\|\mathbb{E}\phi_m(x)\big\|_2^{p^*}\right)^{\frac{1}{p^*}} \;=\; \left(\sum_{m=1}^{M}\big\langle\mathbb{E}\phi_m(x), \mathbb{E}\phi_m(x)\big\rangle^{\frac{p^*}{2}}\right)^{\frac{1}{p^*}} \\ &\overset{\text{Jensen}}{\leq} \left(\sum_{m=1}^{M}\mathbb{E}\big\langle\phi_m(x),\phi_m(x)\big\rangle^{\frac{p^*}{2}}\right)^{\frac{1}{p^*}} \;=\; \sqrt{\Big\|\big(\operatorname{tr}(J_m)\big)_{m=1}^{M}\Big\|_{\frac{p^*}{2}}}\end{aligned} \tag{3}$$

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

$$R_r(H_p) \leq\; \mathbb{E}\sup_{\substack{f_{\boldsymbol{w}}\in H_p,\\ Pf_{\boldsymbol{w}}^2\leq r}}\big\langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^{n}\sigma_i\widetilde{\phi}(x_i)\big\rangle \;+\mathbb{E}\sup_{\substack{f_{\boldsymbol{w}}\in H_p,\\ Pf_{\boldsymbol{w}}^2\leq r}}\big\langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathbb{E}\phi(x)\big\rangle.$$

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

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

Now to bound the second term, we write

$$\mathbb{E}\sup_{\substack{f_{\boldsymbol{w}}\in H_p,\\ Pf_{\boldsymbol{w}}^2\leq r}}\big\langle \boldsymbol{w}, \frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathbb{E}\phi(x)\big\rangle \;\leq\; \sqrt{n}\sup_{\substack{f_{\boldsymbol{w}}\in H_p,\\ Pf_{\boldsymbol{w}}^2\leq r}}\big\langle \boldsymbol{w}, \mathbb{E}\phi(x)\big\rangle.$$

Now observe that we have

$$\big\langle \boldsymbol{w}, \mathbb{E}\phi(x)\big\rangle \overset{\text{Hölder}}{\leq} \|\boldsymbol{w}\|_{2,p}\,\big\|\mathbb{E}\phi(x)\big\|_{2,p^*} \overset{(3)}{\leq} \|\boldsymbol{w}\|_{2,p}\,\sqrt{\Big\|\big(\operatorname{tr}(J_m)\big)_{m=1}^{M}\Big\|_{\frac{p^*}{2}}}$$

as well as $\langle \boldsymbol{w}, \mathbb{E}\phi(x)\rangle = \mathbb{E}f_{\boldsymbol{w}}(x) \leq \sqrt{Pf_{\boldsymbol{w}}^2}$. We finally obtain, putting together the steps above,

$$R_r(H_p) \leq R_r(\widetilde{H}_p) + n^{-\frac{1}{2}}\min\left(\sqrt{r}, D\sqrt{\Big\|\big(\operatorname{tr}(J_m)\big)_{m=1}^{M}\Big\|_{\frac{p^*}{2}}}\right). \tag{4}$$

This shows that there is no loss in working with the centered class instead of the uncentered one.

STEP 2: BOUNDING THE COMPLEXITY OF THE CENTERED CLASS. In this step of the proof we generalize the technique of [19] to multi-kernel classes. First we note that, since the (centered) covariance operator $\mathbb{E}\widetilde{\phi}_m(x) \otimes \widetilde{\phi}_m(x)$ is also a self-adjoint Hilbert-Schmidt operator on $\mathcal{H}_m$, there exists an eigendecomposition $\mathbb{E}\widetilde{\phi}_m(x) \otimes \widetilde{\phi}_m(x) = \sum_{j=1}^{\infty} \widetilde{\lambda}_j^{(m)} \widetilde{\boldsymbol{u}}_j^{(m)} \otimes \widetilde{\boldsymbol{u}}_j^{(m)}$, wherein $(\widetilde{\boldsymbol{u}}_j^{(m)})_{j\geq 1}$ is an orthogonal basis of $\mathcal{H}_m$. Furthermore, the no-correlation assumption (U) entails $\mathbb{E}\widetilde{\phi}_l(x) \otimes \widetilde{\phi}_m(x) = \mathbf{0}$ for all $l \neq m$. As a consequence, for all j and m,

$$P\widetilde{f}_{\boldsymbol{w}}^2 \;=\; \mathbb{E}(\widetilde{f}_{\boldsymbol{w}}(x))^2 \;=\; \mathbb{E}\Big(\sum_{m=1}^{M} \big\langle \boldsymbol{w}_m, \widetilde{\phi}_m(x)\big\rangle\Big)^2 \;=\; \sum_{m=1}^{M}\sum_{j=1}^{\infty} \widetilde{\lambda}_j^{(m)} \Big\langle \boldsymbol{w}_m, \widetilde{\boldsymbol{u}}_j^{(m)}\Big\rangle^2 \tag{5}$$

$$\mathbb{E}\Big\langle \frac{1}{n}\sum_{i=1}^{n} \sigma_i \widetilde{\phi}_m(x_i), \widetilde{\boldsymbol{u}}_j^{(m)}\Big\rangle^2 \;=\; \frac{1}{n}\Big\langle \widetilde{\boldsymbol{u}}_j^{(m)}, \Big(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\,\widetilde{\phi}_m(x_i) \otimes \widetilde{\phi}_m(x_i)\Big)\widetilde{\boldsymbol{u}}_j^{(m)}\Big\rangle \;=\; \frac{\widetilde{\lambda}_j^{(m)}}{n}. \tag{6}$$

Let now $h_1, \ldots, h_M$ be arbitrary nonnegative integers. We can express the LRC in terms of the eigendecompositon as follows

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

so that (5) and (6) yield

$$R_r(\widetilde{H}_p) \overset{(5),\,(6),\,\text{Hölder}}{\leq} \sqrt{\frac{r\sum_{m=1}^{M} h_m}{n}} + D\,\mathbb{E}\Big\|\Big(\sum_{j=h_m+1}^{\infty} \langle \frac{1}{n}\sum_{i=1}^{n} \sigma_i \widetilde{\phi}_m(x_i), \widetilde{\boldsymbol{u}}_j^{(m)}\rangle \widetilde{\boldsymbol{u}}_j^{(m)}\Big)_{m=1}^{M}\Big\|_{2,p^*}.$$

STEP 3: KHINTCHINE-KAHANE'S AND ROSENTHAL'S INEQUALITIES. We use the Khintchine-Kahane (K.-K.) inequality (see Lemma B.2 in the supplementary material) to further bound the right term in the above expression as $E\Big\|\Big(\sum_{j>h_m} \langle \frac{1}{n}\sum_{i=1}^{n} \sigma_i \widetilde{\phi}_m(x_i), \widetilde{\boldsymbol{u}}_j^{(m)}\rangle \widetilde{\boldsymbol{u}}_j^{(m)}\Big)_{m=1}^{M}\Big\|_{2,p^*}$ $\leq \sqrt{\frac{p^*}{n}} \Big(\sum_{m=1}^{M} \mathbb{E}\Big(\sum_{j>h_m} \frac{1}{n}\sum_{i=1}^{n} \langle \widetilde{\phi}_m(x_i), \widetilde{\boldsymbol{u}}_j^{(m)}\rangle^2\Big)^{\frac{p^*}{2}}\Big)^{\frac{1}{p^*}}$. Note that for $p \geq 2$ it holds that $p^*/2 \leq 1$, and thus it suffices to employ Jensen's inequality once again 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 C.1 (Rosenthal + Young; see supplementary material), which yields

$$\begin{aligned}
&\Big(\sum_{m=1}^{M} \mathbb{E}\Big(\sum_{j=h_m+1}^{\infty} \frac{1}{n}\sum_{i=1}^{n} \langle \widetilde{\phi}_m(x_i), \widetilde{\boldsymbol{u}}_j^{(m)}\rangle^2\Big)^{\frac{p^*}{2}}\Big)^{\frac{1}{p^*}} \\
&\qquad \overset{\text{R+Y}}{\leq} \left(\sum_{m=1}^{M} (ep^*)^{\frac{p^*}{2}} \left(\Big(\frac{B}{n}\Big)^{\frac{p^*}{2}} + \Big(\sum_{j=h_m+1}^{\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\langle \widetilde{\phi}_m(x_i), \widetilde{\boldsymbol{u}}_j^{(m)}\rangle^2\Big)^{\frac{p^*}{2}}\right)\right)^{\frac{1}{p^*}} \\
&\qquad \overset{(*)}{\leq} \sqrt{ep^*\left(\frac{BM^{\frac{2}{p^*}}}{n} + \Big(\sum_{m=1}^{M}\Big(\sum_{j=h_m+1}^{\infty} \widetilde{\lambda}_j^{(m)}\Big)^{\frac{p^*}{2}}\Big)^{\frac{2}{p^*}}\right)}
\end{aligned}$$

where for $(*)$ we used the subadditivity of $\sqrt[p^*/2]{\cdot}$. Note that $\forall j, m:\ \widetilde{\lambda}_j^{(m)} \leq \lambda_j^{(m)}$ by the Lidskii-Mirsky-Wielandt theorem since $\mathbb{E}\phi_m(x) \otimes \phi_m(x) = \mathbb{E}\widetilde{\phi}_m(x) \otimes \widetilde{\phi}_m(x) + \mathbb{E}\phi_m(x) \otimes \mathbb{E}\phi_m(x)$. Thus

by the subadditivity of the root function

$$R_r(\widetilde{H}_p) \;\le\; \sqrt{\frac{r\sum_{m=1}^M h_m}{n}} + D\sqrt{\frac{ep^{*2}}{n}\left(\frac{BM^{\frac{2}{p^*}}}{n} + \left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}\right)}$$

$$\le \sqrt{\frac{r\sum_{m=1}^M h_m}{n}} + \sqrt{\frac{ep^{*2}D^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}} + \frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n}. \tag{7}$$

STEP 4: BOUNDING THE COMPLEXITY OF THE ORIGINAL CLASS. Now note that for all nonnegative integers h_m we either have $n^{-\frac{1}{2}}\min\left(\sqrt{r}, D\left(\left\|\left(\operatorname{tr}(J_m)\right)_{m=1}^M\right\|_{\frac{p^*}{2}}\right)^{1/2}\right) \le \left(\frac{ep^{*2}D^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}\right)^{1/2}$ (in case all h_m are zero) or it holds $n^{-\frac{1}{2}}\min\left(\sqrt{r}, D\left(\left\|\left(\operatorname{tr}(J_m)\right)_{m=1}^M\right\|_{\frac{p^*}{2}}\right)^{1/2}\right) \le \left(r\sum_{m=1}^M h_m/n\right)^{1/2}$ (in case that at least one h_m is nonzero) so that in any case we get $n^{-\frac{1}{2}}\min\left(\sqrt{r}, D\left(\left\|\left(\operatorname{tr}(J_m)\right)_{m=1}^M\right\|_{\frac{p^*}{2}}\right)^{1/2}\right) \le \left(\frac{r\sum_{m=1}^M h_m}{n}\right)^{1/2} + \left(\frac{ep^{*2}D^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}\right)^{1/2}$. Thus the following preliminary bound follows from (4) by (7):

$$R_r(H_p) \;\le\; \sqrt{\frac{4r\sum_{m=1}^M h_m}{n}} + \sqrt{\frac{4ep^{*2}D^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}} + \frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n}, \tag{8}$$

for all nonnegative integers $h_m \ge 0$. Later, we will use the above bound (8) for the computation of the excess loss; however, to gain more insight in the bounds' properties, we express it in terms of the truncated spectra of the kernels at the scale r as follows:

STEP 5: RELATING THE BOUND TO THE TRUNCATION OF THE SPECTRA OF THE KERNELS. Next, we 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 $q \ge 1$

$$\sqrt{A_1} + \sqrt{A_2} \;\le\; \sqrt{2(A_1+A_2)} \tag{9}$$

$$\|\boldsymbol{a}_1\|_q + \|\boldsymbol{a}_2\|_q \;\le\; 2^{1-\frac{1}{q}}\|\boldsymbol{a}_1+\boldsymbol{a}_2\|_q \le 2\|\boldsymbol{a}_1+\boldsymbol{a}_2\|_q \tag{10}$$

(the first statement follows from the concavity of the square root function and the second one is readily proved; see Lemma C.3 in the supplementary material) and thus

$$R_r(H_p) \le \sqrt{\frac{16}{n}\left\|\left(rM^{1-\frac{2}{p^*}}h_m + ep^{*2}D^2\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}} + \frac{\sqrt{Be}DM^{\frac{1}{p^*}}p^*}{n},$$

where we used that for all non-negative $\boldsymbol{a} \in \mathbb{R}^M$ and $0 < q < p \le \infty$ it holds

(ℓ_q-to-ℓ_p conversion) $$\|\boldsymbol{a}\|_q = \langle \mathbf{1}, \boldsymbol{a}^q\rangle^{\frac{1}{q}} \overset{\text{Hölder}}{\le} \left(\|\mathbf{1}\|_{(p/q)^*}\|\boldsymbol{a}^q\|_{p/q}\right)^{1/q} = M^{\frac{1}{q}-\frac{1}{p}}\|\boldsymbol{a}\|_p. \tag{11}$$

Since the above holds for all nonnegative integers h_m, the result follows, completing the proof. □

2.1 Lower and Excess Risk Bounds

To investigate the tightness of the presented upper bounds on the LRC of H_p, we consider the case where $\phi_1(x), \dots, \phi_M(x)$ 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 $k_1, \dots, k_M$ are identical and each act on a different coordinate of x.

Theorem 4 (LOWER BOUND)**.** *Assume that the kernels are centered and i.i.d.. Then, there is an absolute constant c such that if $\lambda^{(1)} \ge \frac{1}{nD^2}$ then for all $r \ge \frac{1}{n}$ and $p \ge 1$,*

$$R_r(H_{p,D,M}) \;\ge\; \sqrt{\frac{c}{n}\sum_{j=1}^{\infty}\min(rM, D^2M^{2/p^*}\lambda_j^{(1)})}. \tag{12}$$

Comparing the above lower bound with the upper bounds, we observe that the upper bound of Theorem 2 for centered identical independent kernels is of the order

$O\big(\sqrt{\sum_{j=1}^{\infty} \min\big(rM, D^2 M^{\frac{2}{p^*}} \lambda_j^{(1)}\big)}\big)$, thus matching the rate of the lower bound (the same holds for the bound of Theorem 3). This shows that the upper bounds of the previous section are tight.

As an application of our results to prediction problems such as classification or regression, we also bound the *excess loss* of empirical minimization, $\hat{f} := \operatorname{argmin}_f \frac{1}{n} \sum_{i=1}^n l(f(x_i), y_i)$, w.r.t. to a loss function l: $P(l_{\hat{f}} - l_{f^*}) := \mathbb{E}\, l(\hat{f}(x), y) - \mathbb{E}\, l(f^*(x), y)$, where $f^* := \operatorname{argmin}_f \mathbb{E}\, l(f(x), y)$. We use the analysis of Bartlett et al. [2] to show the following excess risk bound under the assumption of algebraically decreasing eigenvalues of the kernel matrices, i.e. $\exists d > 0, \alpha > 1, \forall m : \lambda_j^{(m)} \leq d\, j^{-\alpha}$ (proof shown in the supplementary material E):

Theorem 5. *Assume that $\|k\|_\infty \leq B$ and $\exists d > 0, \alpha > 1, \forall m : \lambda_j^{(m)} \leq d\, j^{-\alpha}$. Let l be a Lipschitz continuous loss with constant L and assume there is a positive constant F such that $\forall f \in \mathcal{F} : P(f - f^*)^2 \leq F\, P(l_f - l_{f^*})$. Then for all $z > 0$ with probability at least $1 - e^{-z}$ the excess loss of the multi-kernel class H_p can be bounded for $p \in [1, 2]$ as*

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

We see from the above bound that convergence can be almost as slow as $O\big(p^* M^{\frac{1}{p^*}} n^{-\frac{1}{2}}\big)$ (if $\alpha \approx 1$ is small) and almost as fast as $O\big(n^{-1}\big)$ (if α is large).

3 Interpretation of Bounds

In this section, we discuss the rates of Theorem 5 obtained by local analysis bounds, that is

$$\forall t \in [p, 2] : \quad P(l_{\hat{f}} - l_{f^*}) = O\Big(\big(t^* D\big)^{\frac{2}{1+\alpha}} M^{1+\frac{2}{1+\alpha}\left(\frac{1}{t^*}-1\right)} n^{-\frac{\alpha}{1+\alpha}}\Big). \tag{13}$$

On the other hand, the global Rademacher complexity directly leads to a bound of the form [8]

$$\forall t \in [p, 2] : \quad P(l_{\hat{f}} - l_{f^*}) = O\Big(t^* D M^{\frac{1}{t^*}} n^{-\frac{1}{2}}\Big). \tag{14}$$

To compare the above rates, we first assume $p \geq (\log M)^*$ so that the best choice is $t = p$. Clearly, the rate obtained through local analysis is better in n since $\alpha > 1$. Regarding the rate in the number of kernels M and the radius D, a straightforward calculation shows that the local analysis improves over the global one whenever $M^{\frac{1}{p}}/D = O(\sqrt{n})$. Interestingly, this "phase transition" does not depend on α (i.e. the "complexity" of the kernels), but only on p.

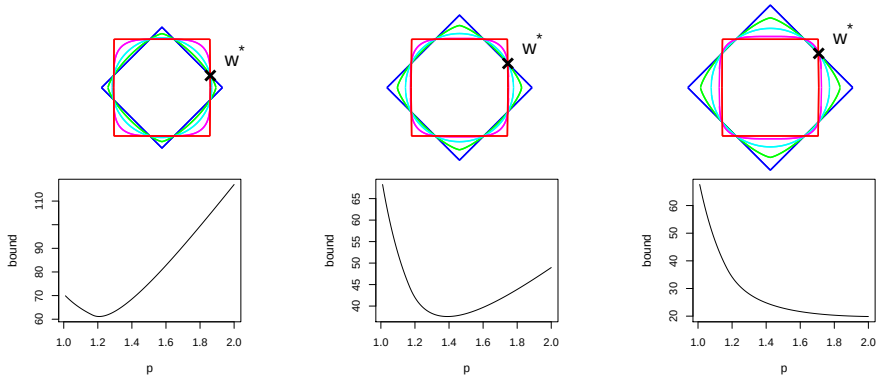
Second, if $p \leq (\log M)^*$, the best choice in (13) and (14) is $t = (\log M)^*$ so that

$$P(l_{\hat{f}} - l_{f^*}) \leq O\Big(\min\Big(Mn^{-1}, \min_{t \in [p,2]} t^* D M^{\frac{1}{t^*}} n^{-\frac{1}{2}}\Big)\Big) \tag{15}$$

and the phase transition occurs for $\frac{M}{D \log M} = O(\sqrt{n})$. Note, that when letting $\alpha \to \infty$ the classical case of aggregation of M basis functions is recovered. This situation is to be compared to the sharp analysis of the optimal convergence rate of convex aggregation of M functions obtained by [27] in the framework of squared error loss regression, which is shown to be $O\Big(\min\Big(\frac{M}{n}, \Big(\frac{1}{n}\log\Big(\frac{M}{\sqrt{n}}\Big)^{1/2}\Big)\Big)\Big)$. This corresponds to the setting studied here with $D = 1, p = 1$ and $\alpha \to \infty$, and we see that our bound recovers (up to log factors) in this case this sharp bound and the related phase transition phenomenon.

Please note that, by introducing an inequality in Eq. (5), Assumption (U)—a similar assumption was also used in [23]—can be relaxed to a more general, RIP-like assumption as used in [16]; this comes at the expense of an additional factor in the bounds (details omitted here).

When Can Learning Kernels Help Performance? As a practical application of the presented bounds, we analyze the impact of the norm-parameter p on the accuracy of ℓ_p-norm MKL in various learning scenarios, showing why an intermediate p often turns out to be optimal in practical applications. As indicated 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 (for example, cf. Figure 1



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

Figure 2: Illustration of the three analyzed learning scenarios (TOP) differing in their soft sparsity of the Bayes hypothesis $\boldsymbol{w}^*$ (parametrized by β) and corresponding values of the bound factor ν_t as a function of p (BOTTOM). A soft sparse (LEFT), a intermediate non-sparse (CENTER), and an almost uniform $\boldsymbol{w}^*$ (RIGHT).

in the introduction). The aim of this section is to relate the theoretical analysis presented here to this empirically observed phenomenon.

To start with, first note that the choice of p only affects the excess risk bound in the factor (cf. Theorem 5 and Equation (13))

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

Let us assume that the Bayes hypothesis can be represented by $\boldsymbol{w}^* \in \mathcal{H}$ such 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 rank-1 kernels. If n is large, the best bias-complexity trade-off 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

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

We can now plot the value ν_p as a function of p fixing α, M, and β. We realized this simulation for $\alpha = 2$, $M = 1000$, and $\beta \in \{0.5, 1, 2\}$.The results are shown in Figure 2. 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 (LEFT) the minimum is attained for a quite small $p = 1.2$, while for the intermediate case (CENTER) $p = 1.4$ is optimal, and finally in the uniformly non-sparse scenario (RIGHT) the choice of $p = 2$ is optimal, i.e. SVM. This means that if the true Bayes hypothesis has an intermediately dense representation (which is frequently encountered in practical applications), our bound gives the strongest generalization guarantees to ℓ_p-norm MKL using an intermediate choice of p.

4 Conclusion

We derived a sharp upper bound on the local Rademacher complexity of ℓ_p-norm multiple kernel learning. 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.

Acknowledgments

We thank Peter L. Bartlett and K.-R. Müller for valuable comments. This work was supported by the German Science Foundation (DFG MU 987/6-1, RA 1894/1-1) and by the European Community's 7th Framework Programme under the PASCAL2 Network of Excellence (ICT-216886) and under the E.U. grant agreement 247022 (MASH Project).

References

[1] 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.

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

[3] 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.

[4] C. Campbell and Y. Ying. *Learning with Support Vector Machines*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.

[5] 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. Video `http://videolectures.net/icml09_cortes_clkh/`.

[6] 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://videolectures.net/lkasok08_whistler/`, Video `http://www.cs.nyu.edu/learning_kernels`.

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

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

[9] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[10] 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.

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

[12] J.-P. Kahane. *Some random series of functions*. Cambridge University Press, 2nd edition, 1985.

[13] 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.

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

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

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

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

[18] 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.

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

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

[21] 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.

[22] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[23] 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.

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

[25] J. R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(03):417–424, 1980.

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

[27] 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.

Supplementary Material

A Details of the Experiment Presented in the Introduction

We obtained the *dingshen* data set including the training and test split used in [4]. The dingshen data set consists of 27 fold classes with 313 proteins used for training and 385 for testing. There are a number of observational features relevant to predicting fold class, and in this study, 12 different informative data-types were used. This included the RNA sequence and various physical measurements such as hydrophobicity, polarity and van der Waals volume resulting in 12 kernels [4].

We precisely replicate the experimental setup of [4]: we carry out MKL via one-vs.-rest SVMs to deal with the multiple classes and report on test set accuracy. However, in contrast to [4], we investigate $\ell_{p>1}$-norm MKL instead of just ℓ_1-norm MKL. We perform model selection by cross validation on the training set over $C \in 10^{[-4,-3.5,\ldots,4]}$.

Results The results are shown in Figure 1 (LEFT) in the introduction of this paper. The vertical bars indicate the test set accuracy for the single-kernel SVMs (e.g., H denotes the Hydrophobicity kernel, P the Polarity kernel, etc.). The horizontal bar indicates the performance of the MKL algorithm with all data-types included. The best single-kernel SVM is the one using the SW2-kernel and has a test set accuracy of 64.0%; in contrast, the SVM using a uniform kernel combintation achieves a substantially better accuracy of 68.9%, which is slightly better than the 68.4% that ℓ_1-norm MKL achieves. Interestingly, there is a huge improvement in using non-sparse $\ell_{p>1}$-norm MKL: the best performing norm is $p = 1.14$, which has an impressive accuracy of 74.4%. This indicates the relevance of this method for the application domain.

Figure 1 (RIGHT) gives the values of the kernel coefficients $\boldsymbol{\theta}$. We observe that ℓ_1-norm MKL puts most of the weights into SW1- and SW2-kernels, which also have the highest single-kernel performance. Generally, the chosen kernel combinations nicely reflect the single-kernel performances as determined by the single-kernel SVMs. The $\ell_{p>1}$-norm variants yield precisely the same "ranking" of weights θ_i but stronger distributes the weights among the kernels.

Interpretation The superior performance of $\ell_{1.14}$-norm MKL compared to ℓ_1-norm MKL and the SVM using a uniform kernel combination indicates that all 12 types of data are relevant—but not equally relevant at all. For example, the features SW1 and SW2, which are based on sequence alignments, appear to be more informative than the others.

To further analyze the result, we compute the pairwise kernel alignments shown in Figure A.1. One can see from the figure that the Kernels L1–L30 and SW1–SW2 corelate quite strongly. This resembles the similarity in the generation process of those kernels (they differ by different parameter values). However, the other kernels correlate surprisingly few—this indicates that here orthogonal information is contained in the kernels. Therefore discarding or overly downgrading one of those kernels can be disadvantageous, which explains the poor ℓ_1-norm MKL performance. On the other hand we know that from the single-kernel performances that not all kernels are equally informative, which explains the rather bad performance of the uniform-combination SVM. We conclude that an intermediate norms must be optimal—and this also what we observe in terms of test errors.

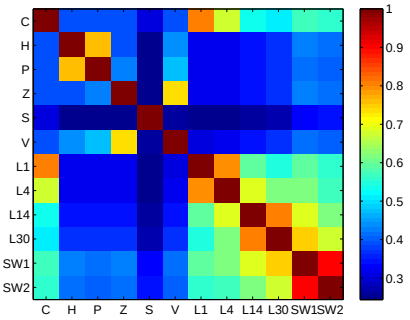


Figure A.1: Pairwise kernel alignments of the protein fold prediction experiment.

B Global Rademacher Complexity Bound

Proof of Proposition 1 (GRC Upper Bound). First note that it suffices to prove the result for $t = p$ as trivially $\|\boldsymbol{w}\|_{2,t} \leq \|\boldsymbol{w}\|_{2,p}$ holds for all $t \geq p$ so that $H_p \subseteq H_t$ and therefore $R(H_p) \leq R(H_t)$. We can use a block-structured version of Hölder's inequality (cf. Lemma B.1) and the Khintchine-Kahane (K.-K.) inequality (cf. Lemma B.2) to bound the empirical version of the global RC as follows:

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

what was to show. □

The following result gives a block-structured version of Hölder's inequality

Lemma B.1 (BLOCK-STRUCTURED HÖLDER INEQUALITY)**.** *Let $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$, $\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m) \in \mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_M$. Then, for any $p \geq 1$, it holds*

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

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

$$\begin{aligned}
\langle \boldsymbol{v}, \boldsymbol{w} \rangle &= \sum_{m=1}^{M} \langle \boldsymbol{v}_m, \boldsymbol{w}_m \rangle \overset{\text{C.-S.}}{\leq} \sum_{m=1}^{M} \|\boldsymbol{v}\|_2 \|\boldsymbol{w}\|_2 \\
&= \big\langle (\|\boldsymbol{v}_1\|_2, \ldots, \|\boldsymbol{v}_M\|_2), (\|\boldsymbol{w}_1\|_2, \ldots, \|\boldsymbol{w}_M\|_2) \big\rangle. \\
&\overset{\text{Hölder}}{\leq} \|\boldsymbol{v}\|_{2,p} \|\boldsymbol{w}\|_{2,p^*}
\end{aligned}$$

□

The following inequality is known as the Khintchine-Kahane inequality [12]; we employ the constants taken from Lemma 3.3.1 and Proposition 3.4.1 in [17]:

Lemma B.2 (KHINTCHINE-KAHANE INEQUALITY)**.** *Let be $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M \in \mathcal{H}$. Then, for any $p \geq 1$, it holds $\mathbb{E}_{\boldsymbol{\sigma}} \big\| \sum_{i=1}^{n} \sigma_i \boldsymbol{v}_i \big\|_2^p \leq \Big(c \sum_{i=1}^{n} \big\|\boldsymbol{v}_i\big\|_2^2 \Big)^{\frac{p}{2}}$, where $c = \max(1, p^* - 1)$. In particular the result holds for $c = p^*$.*

C Local Rademacher Complexity Bound

Proof of Theorem 3 (LRC Upper Bound, $p > 2$). The eigendecomposition $\mathbb{E}\phi(x) \otimes \phi(x) = \sum_{j=1}^{\infty} \lambda_j \boldsymbol{u}_j \otimes \boldsymbol{u}_j$ yields

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

and, for all j

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

Therefore, we can use, for any nonnegative integer h, the Cauchy-Schwarz inequality and a block-structured version of Hölder's inequality (see Lemma B.1) to bound the local Rademacher complexity as follows:

$$\begin{aligned}
R_r(H_p) &= \mathbb{E} \sup_{f_{\boldsymbol{w}} \in H_p : P f_{\boldsymbol{w}}^2 \leq r} \big\langle \boldsymbol{w}, \frac{1}{n} \sum_{i=1}^n \sigma_i \phi(x)_i \big\rangle \\
&= \mathbb{E} \sup_{f_{\boldsymbol{w}} \in H_p : P f_{\boldsymbol{w}}^2 \leq r} \Big\langle \sum_{j=1}^h \sqrt{\lambda_j} \langle \boldsymbol{w}, \boldsymbol{u}_j \rangle \boldsymbol{u}_j, \sum_{j=1}^h \sqrt{\lambda_j}^{-1} \langle \frac{1}{n} \sum_{i=1}^n \sigma_i \phi(x)_i, \boldsymbol{u}_j \rangle \boldsymbol{u}_j \Big\rangle \\
&\qquad + \big\langle \boldsymbol{w}, \sum_{j=h+1}^\infty \langle \frac{1}{n} \sum_{i=1}^n \sigma_i \phi(x)_i, \boldsymbol{u}_j \rangle \boldsymbol{u}_j \big\rangle \\
&\overset{\text{C.-S., (C.1), (C.2)}}{\leq} \sqrt{\frac{rh}{n}} + \mathbb{E} \sup_{f_{\boldsymbol{w}} \in H_p} \big\langle \boldsymbol{w}, \sum_{j=h+1}^\infty \langle \frac{1}{n} \sum_{i=1}^n \sigma_i \phi(x)_i, \boldsymbol{u}_j \rangle \boldsymbol{u}_j \big\rangle \\
&\overset{\text{Hölder}}{\leq} \sqrt{\frac{rh}{n}} + D \mathbb{E} \Big\| \sum_{j=h+1}^\infty \langle \frac{1}{n} \sum_{i=1}^n \sigma_i \phi(x)_i, \boldsymbol{u}_j \rangle \boldsymbol{u}_j \Big\|_{2,p^*} \\
&\overset{\ell_{\frac{p^*}{2}}\text{-to-}\ell_2}{\leq} \sqrt{\frac{rh}{n}} + D M^{\frac{1}{p^*} - \frac{1}{2}} \mathbb{E} \Big\| \sum_{j=h+1}^\infty \langle \frac{1}{n} \sum_{i=1}^n \sigma_i \phi(x)_i, \boldsymbol{u}_j \rangle \boldsymbol{u}_j \Big\|_2 \\
&\overset{\text{Jensen}}{\leq} \sqrt{\frac{rh}{n}} + D M^{\frac{1}{p^*} - \frac{1}{2}} \Bigg(\sum_{j=h+1}^\infty \underbrace{\mathbb{E} \langle \frac{1}{n} \sum_{i=1}^n \sigma_i \phi(x)_i, \boldsymbol{u}_j \rangle^2}_{\overset{\text{(C.2)}}{\leq} \frac{\lambda_j}{n}} \Bigg)^{\frac{1}{2}} \\
&\leq \sqrt{\frac{rh}{n}} + \sqrt{\frac{D^2 M^{\frac{2}{p^*} - 1}}{n} \sum_{j=h+1}^\infty \lambda_j}.
\end{aligned}$$

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

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

□

Lemma C.1 (ROSENTHAL + YOUNG)**.** *Let $X_1, \ldots, X_n$ be independent nonnegative random variables satisfying $\forall i : X_i \leq B < \infty$ almost surely. Then, denoting $c_q = (2qe)^q$, for any $q \geq \frac{1}{2}$ it holds*

$$\mathbb{E} \Big(\frac{1}{n} \sum_{i=1}^n X_i \Big)^q \leq c_q \Big(\Big(\frac{B}{n} \Big)^q + \Big(\frac{1}{n} \sum_{i=1}^n \mathbb{E} X_i \Big)^q \Big).$$

Proof. 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 to the sequence $X_1, \ldots, X_n$ thereby using the optimal constants computed in [11], 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

$$\mathbb{E} \Big(\frac{1}{n} \sum_{i=1}^n X_i \Big)^q \leq c_q \max \left(\frac{1}{n^q} \sum_{i=1}^n \mathbb{E} X_i^q, \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^q \right). \tag{C.3}$$

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:

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

It thus follows from (C.3) 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 C.2 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 C.2. *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 $q \geq 1$:*

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

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

$$\begin{aligned}\mathbb{E}^q &= \frac{1}{e}\left(0 + \sum_{k=1}^q \frac{k^q}{k!} + \sum_{k=q+1}^\infty \frac{k^q}{k!}\right) \\ &= \frac{1}{e}\left(\sum_{k=1}^q \frac{k^{q-1}}{(k-1)!} + \sum_{k=q+1}^\infty \frac{k^q}{k!}\right) \\ &\leq \frac{1}{e}\left(q^q + \sum_{k=q+1}^\infty \frac{k^q}{k!}\right) \qquad \text{(C.4)}\end{aligned}$$

(C.5)

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 (C.4)

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

which was to show. □

Lemma C.3. *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

$$\|\boldsymbol{a}\|_q^q + \|\boldsymbol{b}\|_q^q \leq \|\boldsymbol{a}+\boldsymbol{b}\|_q^q . \tag{C.6}$$

We conclude by ℓ_q-to-ℓ_1 conversion (see (11))

$$\begin{aligned} \|\boldsymbol{a}\|_q + \|\boldsymbol{b}\|_q &= \left\|\left(\|\boldsymbol{a}\|_q, \|\boldsymbol{b}\|_q\right)\right\|_1 \overset{(11)}{\leq} 2^{1-\frac{1}{q}} \left\|\left(\|\boldsymbol{a}\|_q, \|\boldsymbol{b}\|_q\right)\right\|_q \\ &= 2^{1-\frac{1}{q}}\left(\|\boldsymbol{a}\|_q^q + \|\boldsymbol{b}\|_q^q\right)^{\frac{1}{q}} \overset{(C.6)}{\leq} 2^{1-\frac{1}{q}} \|\boldsymbol{a}+\boldsymbol{b}\|_q , \end{aligned}$$

which completes the proof. □

D LRC Lower Bound

Proof of Theorem 4 (LRC Lower Bound). First note that since the $\phi_i(x)$ are centered and uncorrelated, that

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

Now it follows

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

so that we can use the i.i.d. assumption on $\phi_m(x)$ to equivalently rewrite the last term as

$$\begin{aligned}
R_r(H_{p,D,M}) &\overset{(\phi_m(x))_{1\le m\le M}\text{ i.i.d.}}{\geq} \mathbb{E} \sup_{\substack{\boldsymbol{w}^{(1)}:\ \langle \boldsymbol{w}^{(1)},\phi_1(x)\rangle^2 \le r/M \\ \|\boldsymbol{w}^{(1)}\|_2 \le DM^{-\frac{1}{p}}}} \left\langle M\boldsymbol{w}^{(1)}, \frac{1}{n}\sum_{i=1}^n \sigma_i\phi_1(x_i)\right\rangle \\
&= \mathbb{E} \sup_{\substack{\boldsymbol{w}^{(1)}:\ \langle M\boldsymbol{w}^{(1)},\phi_1(x)\rangle^2 \le rM \\ \|M\boldsymbol{w}^{(1)}\|_2 \le DM^{\frac{1}{p^*}}}} \left\langle M\boldsymbol{w}^{(1)}, \frac{1}{n}\sum_{i=1}^n \sigma_i\phi_1(x_i)\right\rangle \\
&= \mathbb{E} \sup_{\substack{\boldsymbol{w}^{(1)}:\ \langle \boldsymbol{w}^{(1)},\phi_1(x)\rangle^2 \le rM \\ \|\boldsymbol{w}^{(1)}\|_2 \le DM^{\frac{1}{p^*}}}} \left\langle \boldsymbol{w}^{(1)}, \frac{1}{n}\sum_{i=1}^n \sigma_i\phi_1(x_i)\right\rangle \\
&= R_{rM}(H_{1,DM^{1/p^*},1})
\end{aligned}$$

□

In [19] 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 the lower bound of Theorem 4.

E Excess Risk Bound

In [2, 15] it was shown that the rate of convergence of the excess risk is basically determined by the fixed point of the local Rademacher complexity. To this end we show:

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

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

Proof. For this proof we make use of the bound (8) on the local Rademacher complexity. Defining $a = \frac{4F^2\sum_{m=1}^M h_m}{n}$ and $b = 4FL\sqrt{\frac{ep^{*2}D^2}{n}\left\|\left(\sum_{j=h_m+1}^\infty \lambda_j^{(m)}\right)_{m=1}^M\right\|_{\frac{p^*}{2}}} + \frac{2\sqrt{Be}DFLM^{\frac{1}{p^*}}p^*}{n}$, in order to find a fixed point of (8) 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 $r^* \geq a + 2b$. 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 for all m there exist $d_m > 0$ and $\alpha_m > 1$ such that $\lambda_j^{(m)} \leq d_m j^{-\alpha_m}$. This is a common assumption and, for example, met for finite rank kernels and convolution kernels. We are now ready to prove Theorem 5.

Proof of Theorem 5 (Excess Risk Bound). First note that

$$\sum_{j>h_m} \lambda_j^{(m)} \leq d_m \sum_{j>h_m} j^{-\alpha_m} \leq d_m \int_{h_m}^\infty x^{-\alpha_m}dx = d_m\Big[\frac{1}{1-\alpha_m}x^{1-\alpha_m}\Big]_{h_m}^\infty = -\frac{d_m}{1-\alpha_m}h_m^{1-\alpha_m}. \tag{E.1}$$

To exploit the above fact (E.1), first note that by ℓ_p-to-ℓ_q conversion

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

so that we can translate the result of the previous lemma by (9), (10), and (11) into

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

Inserting the result of (E.1) into the above bound and setting the derivative with respect to h_m to zero we find the optimal h_m as

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

Resubstituting the above into (E.2) 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).

Therefore, denoting $d := \max_{m\in\{1,\ldots,M\}} d_m$ and $\alpha := \min_{m\in\{1,\ldots,M\}}\alpha_m$, and $h_{\max} := \left(4dep^{*2}D^2F^{-2}L^2M^{\frac{2}{p^*}-2}n\right)^{\frac{1}{1+\alpha_{\min}}}$, we can upper-bound (E.2) by

$$\begin{aligned} r^* &\leq 8F\sqrt{\frac{3-\alpha}{1-\alpha}F^2M^2h_{\max}^2n^{-2}} + \frac{4\sqrt{Be}DFLM^{\frac{1}{p^*}}p^*}{n} \\ &\leq 8\sqrt{\frac{3-\alpha}{1-\alpha}}F^2Mh_{\max}n^{-1} + \frac{4\sqrt{Be}DFLM^{\frac{1}{p^*}}p^*}{n} \\ &\leq 16\sqrt{e\frac{3-\alpha}{1-\alpha}}(dD^2L^2p^{*2})^{\frac{1}{1+\alpha}}F^{\frac{2\alpha}{1+\alpha}}M^{1+\frac{2}{1+\alpha}\left(\frac{1}{p^*}-1\right)}n^{-\frac{\alpha}{1+\alpha}} \\ &\quad + \frac{4\sqrt{Be}DFLM^{\frac{1}{p^*}}p^*}{n}\,. \end{aligned} \tag{E.3}$$

We have thus proved the theorem, which follows by the above inequality, Lemma E.2, and the fact that our class H_p ranges in $BDM^{\frac{1}{p^*}}$. □

The above proof uses the following result, which is a slight modification of Corollary 5.3 in [2] that is well-tailored to the class studied in this paper.[1]

Lemma E.2 (BARTLETT, BOUSQUET, AND MENDELSON, 2005 [2]). *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 $\forall f\in\mathcal{F}:\ P(f-f^*)^2 \leq F\,P(l_f - l_{f^*})$. Then, denoting by r^* the fixed point of*

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

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

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

[1] We exploit the improved constants from Theorem 3.3 in [2] because an absolute convex class is star-shaped. Compared to Corollary 5.3 in [2] 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.