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.

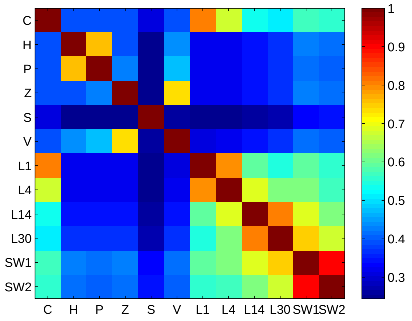

Figure A.1: Pairwise kernel alignments of the protein fold prediction experiment.

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:

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

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 $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$:

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

☐

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

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

and, for all j

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

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:

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

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

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

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 $q \geq \frac{1}{2}$

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

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 $e^{\tau_k} k \le e^{\tau_{k+q}}(k+q)$ can be shown by some algebra using $\frac{1}{12k+1} < \tau_k < \frac{1}{12k}$. Now by (C.4)

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

which was to show. □

Lemma C.3. *For any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}_+^m$ it holds for all $q \ge 1$*

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

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

$$\forall i = 1, \ldots, m: \; a_i^q + b_i^q \le \big(a_i + b_i\big)^q$$

and thus

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

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

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

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

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

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

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

□

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*

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

***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

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

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

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

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 $d := \max_{m\in\{1,\dots,M\}} d_m$ and $\alpha := \min_{m\in\{1,\dots,M\}}\alpha_m$, 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

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

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.