Generalization Bounds for Learning Kernels

Corinna Cortes CORINNA@GOOGLE.COM
Google Research, 76 Ninth Avenue, New York, NY 10011.

Mehryar Mohri MOHRI@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences and Google Research, 251 Mercer Street, New York, NY 10012.

Afshin Rostamizadeh ROSTAMI@CS.NYU.EDU
Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012.

Abstract

This paper presents several novel generalization bounds for the problem of learning kernels based on a combinatorial analysis of the Rademacher complexity of the corresponding hypothesis sets. Our bound for learning kernels with a convex combination of p base kernels using L_1 regularization admits only a $\sqrt{\log p}$ dependency on the number of kernels, which is *tight* and considerably more favorable than the previous best bound given for the same problem. We also give a novel bound for learning with a non-negative combination of p base kernels with an L_2 regularization whose dependency on p is also *tight* and only in $p^{1/4}$. We present similar results for L_q regularization with other values of q, and outline the relevance of our proof techniques to the analysis of the complexity of the class of linear functions. Experiments with a large number of kernels further validate the behavior of the generalization error as a function of p predicted by our bounds.

1. Introduction

Kernel methods are widely used in statistical learning (Schölkopf & Smola, 2002; Shawe-Taylor & Cristianini, 2004). Positive definite symmetric (PDS) kernels implicitly specify an inner product in a Hilbert space where large-margin techniques are used for learning and estimation. They can be combined with algorithms such as support vector machines (SVMs) (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1998) or other kernel-based algorithms to form powerful learning techniques.

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010.

But the choice of the kernel, which is critical to the success of these algorithms, is typically left to the user. Rather than requesting the user to commit to a specific kernel, which may not be optimal, especially if the user's prior knowledge about the task is poor, learning kernel methods require the user only to supply a family of kernels. The learning algorithm then selects both the specific kernel out of that family, and the hypothesis defined based on that kernel.

There is a large body of literature dealing with various aspects of the problem of learning kernels, including theoretical questions, optimization problems related to this problem, and experimental results (Lanckriet et al., 2004; Argyriou et al., 2005; 2006; Srebro & Ben-David, 2006; Lewis et al., 2006; Zien & Ong, 2007; Bach, 2008; Cortes et al., 2009a; Ying & Campbell, 2009). Some of this previous work considers families of Gaussian kernels (Micchelli & Pontil, 2005) or hyperkernels (Ong et al., 2005). Non-linear combinations of kernels have also been recently considered by Bach (2008) and Cortes et al. (2009b). But, the most common family of kernels examined is that of non-negative or convex combinations of some fixed kernels constrained by a trace condition, which can be viewed as an L_1 regularization (Lanckriet et al., 2004), or by an L_2 regularization (Cortes et al., 2009a).

This paper presents several novel generalization bounds for the problem of learning kernels with the family of non-negative combinations of base kernels with an L_1 or L_2 constraint, or L_q constraints with some other values of q. One of the first learning bounds given by Lanckriet et al. (2004) for the family of convex combinations of p base kernels with an L_1 constraint has the following form: $R(h) \leq \widehat{R}_\rho(h) + O\big(\frac{1}{\sqrt{m}} \sqrt{\max_{k=1}^p \operatorname{Tr}(\mathbf{K}_k) \max_{i=1}^p (\|\mathbf{K}_k\| / \operatorname{Tr}(\mathbf{K}_k))/\rho^2}\big)$, where $R(h)$ is the generalization error of a hypothesis h, $\widehat{R}_\rho(h)$ is the fraction of training points with margin less than ρ, and $\mathbf{K}_k$ is the kernel matrix associated

to the kth base kernel. This bound and a similar one by Bousquet & Herrmann (2002) were both shown by Srebro & Ben-David (2006) to be always larger than one. Another bound by Lanckriet et al. (2004) for the family of linear or non-convex combinations of kernels was also shown, by the same authors, to be always larger than one.

But Lanckriet et al. (2004) also presented a multiplicative bound for convex combinations of base kernels with an L_1 constraint that is of the form $R(h) \leq \widehat{R}_\rho(h) + O\Big(\sqrt{\frac{p/\rho^2}{m}}\Big)$. This bound converges and can perhaps be viewed as the first informative generalization bound for this family of kernels. However, the dependence of this bound on the number of kernels p is multiplicative which therefore does not encourage the use of too many base kernels. Srebro & Ben-David (2006) presented a generalization bound based on the pseudo-dimension of the family of kernels that significantly improved on this bound. Their bound has the form $R(h) \leq \widehat{R}_\rho(h) + \widetilde{O}\Big(\sqrt{\frac{p+R^2/\rho^2}{m}}\Big)$, where the notation $\widetilde{O}(\cdot)$ hides logarithmic terms and where R^2 is an upper bound on $K_k(x,x)$ for all points x and base kernels K_k, $k \in [1,p]$. Thus, disregarding logarithmic terms, their bound is only additive in p. Their analysis also applies to other families of kernels. Ying & Campbell (2009) also gave generalization bounds for learning kernels based on the notion of Rademacher chaos complexity and the pseudo-dimension of the family of kernels used. For a pseudo-dimension of p as in the case of a convex combination of p base kernels, their bound is in $O(\sqrt{p\,(R^2/\rho^2)(\log(m)/m)})$ and is thus multiplicative in p. It seems to be weaker than the bound of Lanckriet et al. (2004) and that of Srebro & Ben-David (2006) for such kernel families.

We present new generalization bounds for the family of convex combinations of base kernels and an L_1 constraint that have only a logarithmic dependency on p. Our learning bounds are based on a combinatorial analysis of the Rademacher complexity of the hypothesis set considered and have the form: $R(h) \leq \widehat{R}_\rho(h) + O\Big(\sqrt{\frac{(\log p)R^2/\rho^2}{m}}\Big)$. Our bound is simpler, contains no other extra logarithmic term, and its $\sqrt{\log p}$ dependency is *tight*. Thus, this represents a substantial improvement over the previous best bounds for this problem. Our bound is also valid for a very large number of kernels, in particular for $p \gg m$, while the previous bounds were not informative in that case.

We note that Koltchinskii & Yuan (2008) also presented a bound with logarithmic dependence on p in the context of the study of large ensembles of kernel machines. However, their analysis is specific to the family of kernel-based regularization algorithms and requires the loss function to be strongly convex, which rules out for example the binary classification loss function. Also, both the statement of the result and the proof seem to be considerably more complicated than ours.

We also give a novel bound for learning with a non-negative combination of p base kernels with an L_2 regularization whose dependency on p is also *tight* and only in $p^{1/4}$. We present similar results for L_q regularization with other values of q.

The next section (Section 2) defines the family of kernels and hypothesis sets we examine. Section 3 presents a bound on the Rademacher complexity of the class of convex combinations of base kernels with an L_1 constraint and a generalization bound for binary classification directly derived from that result. Similarly, Section 4 presents first a bound on the Rademacher complexity, then a generalization bound for L_q regularization for some other values of $q > 1$. We make a number of comparisons with existing bounds and conclude by discussing the relevance of our proof techniques to the analysis of the complexity of the class of linear functions (Section 5).

2. Preliminaries

Let $\mathcal{X}$ denote the input space. For any kernel function K, we denote by $\mathbf{\Phi}_K \colon x \mapsto \mathbb{H}_K$ the feature mapping from $\mathcal{X}$ to the reproducing kernel Hilbert space $\mathbb{H}_K$ induced by K. Most learning kernel algorithms are based on a hypothesis H_p^q set derived from a non-negative combinations of a fixed set of $p \geq 1$ kernels $K_1, \ldots, K_p$ with the mixture weights obeying an L_q constraint:

$$H_p^q = \big\{x \mapsto \mathbf{w} \cdot \mathbf{\Phi}_K(x) \colon K = \sum_{k=1}^{p} \mu_k K_k, \boldsymbol{\mu} \in \Delta_q, \|\mathbf{w}\| \leq 1\big\},$$

with $\Delta_q = \{\boldsymbol{\mu} \colon \boldsymbol{\mu} \geq 0, \sum_{k=1}^{p} \mu_k^q = 1\}$. Linear combinations with possibly negative mixture weights have also been considered in the literature, e.g., (Lanckriet et al., 2004), with the additional requirement that the combined kernel be PDS.

We bound, for different values of q, including $q = 1$ and $q = 2$, the empirical Rademacher complexity $\widehat{\mathfrak{R}}_S(H_p^q)$ of these families for an arbitrary sample S of size m, which immediately yields a generalization bound for learning kernels based on these families of hypotheses.

For a fixed sample $S = (x_1, \ldots, x_m)$, the empirical Rademacher complexity of H is defined by

$$\widehat{\mathfrak{R}}_S(H) = \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \Big],$$

where the expectation is taken over $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)^\top$ where $\sigma_i \in \{-1, +1\}$, $i \in [1,m]$, are independent uniform random variables.

For any kernel function K, we denote by $\mathbf{K} = [K(x_i, x_j)] \in \mathbb{R}^{m\times m}$ its kernel matrix associated to the sample S. Let $\mathbf{w}_S = \sum_{i=1}^m \alpha_i \mathbf{\Phi}_K(x_i)$ be the orthogonal projection of $\mathbf{w}$ on $\mathbb{H}_S = \mathrm{span}(\mathbf{\Phi}_K(x_1), \ldots, \mathbf{\Phi}_K(x_m))$. Then, $\mathbf{w}$ can be written as $\mathbf{w} = \mathbf{w}_S + \mathbf{w}^\perp$, with $\mathbf{w}_S \cdot \mathbf{w}^\perp = 0$. Thus, $\|\mathbf{w}\|^2 = \|\mathbf{w}_S\|^2 + \|\mathbf{w}^\perp\|^2$, which, in view of $\|\mathbf{w}\| \leq 1$ implies $\|\mathbf{w}_S\|^2 \leq 1$. Since $\|\mathbf{w}_S\|^2 = \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$, this implies

$$\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1. \tag{1}$$

Observe also that for all $x \in S$,

$$h(x) = \mathbf{w} \cdot \mathbf{\Phi}_K(x) = \mathbf{w}_S \cdot \mathbf{\Phi}_K(x) = \sum_{i=1}^m \alpha_i K(x_i, x). \tag{2}$$

Conversely, any function $\sum_{i=1}^m \alpha_i K(x_i, \cdot)$ with $\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1$ is clearly an element of H_p^1.

Proposition 1. *Let $q, r \geq 1$ with $\frac{1}{q} + \frac{1}{r} = 1$. For any sample S of size m, the empirical Rademacher complexity of the hypothesis set H_p^q can be expressed as*

$$\widehat{\mathfrak{R}}_S(H_p^q) = \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \left[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_r}\right]$$

with $\mathbf{u}_{\boldsymbol{\sigma}} = (\boldsymbol{\sigma}^\top \mathbf{K}_1 \boldsymbol{\sigma}, \ldots, \boldsymbol{\sigma}^\top \mathbf{K}_p \boldsymbol{\sigma})^\top$.

Proof. Fix a sample $S = (x_1, \ldots, x_m)$, and denote by $\mathcal{M}_q = \{\boldsymbol{\mu} \geq 0 \colon \|\boldsymbol{\mu}\|_q = 1\}$ and by $\mathcal{A} = \{\boldsymbol{\alpha} \colon \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1\}$. Then, in view of (1) and (2), the Rademacher complexity $\widehat{\mathfrak{R}}_S(H_p^q)$ can be expressed as follows:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(H_p^q) &= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[\sup_{h \in H_p^q} \sum_{i=1}^m \sigma_i h(x_i) \Big] \\
&= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[\sup_{\boldsymbol{\mu} \in \mathcal{M}_q, \boldsymbol{\alpha} \in \mathcal{A}} \sum_{i,j=1}^m \sigma_i \alpha_j K(x_i, x_j) \Big] \\
&= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[\sup_{\boldsymbol{\mu} \in \mathcal{M}_q, \boldsymbol{\alpha} \in \mathcal{A}} \boldsymbol{\sigma}^\top \mathbf{K} \boldsymbol{\alpha} \Big].
\end{aligned}$$

Now, by the Cauchy-Schwarz inequality, the supremum $\sup_{\boldsymbol{\alpha} \in \mathcal{A}} \boldsymbol{\sigma}^\top \mathbf{K} \boldsymbol{\alpha}$ is reached for $\mathbf{K}^{1/2} \boldsymbol{\alpha}$ collinear with $\mathbf{K}^{1/2} \boldsymbol{\sigma}$, which gives $\sup_{\boldsymbol{\alpha} \in \mathcal{A}} \boldsymbol{\sigma}^\top \mathbf{K} \boldsymbol{\alpha} = \sqrt{\boldsymbol{\sigma}^\top \mathbf{K} \boldsymbol{\sigma}}$. Thus,

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(H_p^q) &= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[\sup_{\boldsymbol{\mu} \in \mathcal{M}_q} \sqrt{\boldsymbol{\sigma}^\top \mathbf{K} \boldsymbol{\sigma}} \Big] \\
&= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[\sup_{\boldsymbol{\mu} \in \mathcal{M}_q} \sqrt{\boldsymbol{\mu} \cdot \mathbf{u}_{\boldsymbol{\sigma}}} \Big].
\end{aligned}$$

By the definition of the dual norm, $\sup_{\boldsymbol{\mu} \in \mathcal{M}_q} \boldsymbol{\mu} \cdot \mathbf{u}_{\boldsymbol{\sigma}} = \|\mathbf{u}_{\boldsymbol{\sigma}}\|_r$, which gives $\widehat{\mathfrak{R}}_S(H_p^q) = \frac{1}{m} \mathrm{E}_{\boldsymbol{\sigma}} \left[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_r}\right]$. □

3. Rademacher complexity bound for H_p^1

Our bounds on the empirical Rademacher complexity of the families H_p^1 or H_p^q for $q = 2$ or other values of q relies on the following result, which we prove using a combinatorial argument (see appendix).

Lemma 1. *Let $\mathbf{K}$ be the kernel matrix of a kernel function K associated to a sample S. Then, for any integer r, the following inequality holds:*

$$\mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[(\boldsymbol{\sigma}^\top \mathbf{K} \boldsymbol{\sigma})^r \Big] \leq \Big(\eta_0 r \operatorname{Tr}[\mathbf{K}] \Big)^r,$$

where $\eta_0 = \frac{23}{22}$.

This result can be viewed as a Khintchine-Kahane type inequality. In fact, it might be possible to benefit from the best constants for the vectorial version of this inequality to further improve the constant of the lemma. We will discuss this connection and its benefits in a longer version of this paper. For $r = 1$, the result holds with η_0 replaced with 1 as seen in classical derivations for the estimation of the Rademacher complexity of linear classes.

Theorem 1. *For any sample S of size m, the empirical Rademacher complexity of the hypothesis set H_p^1 can be bounded as follows:*

$$\forall r \in \mathbb{N}, r \geq 1, \quad \widehat{\mathfrak{R}}_S(H_p^1) \leq \frac{\sqrt{\eta_0 r \|\mathbf{u}\|_r}}{m},$$

where $\mathbf{u} = (\operatorname{Tr}[\mathbf{K}_1], \ldots, \operatorname{Tr}[\mathbf{K}_p])^\top$ and $\eta_0 = \frac{23}{22}$.

Proof. By Proposition 1, $\widehat{\mathfrak{R}}_S(H_p^1) = \frac{1}{m} \mathrm{E}_{\boldsymbol{\sigma}} \left[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_\infty}\right]$. Since for any $r \geq 1$, $\|\mathbf{u}_{\boldsymbol{\sigma}}\|_\infty \leq \|\mathbf{u}_{\boldsymbol{\sigma}}\|_r$, we can upper bound the Rademacher complexity as follows:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(H_p^1) &\leq \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \left[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_r}\right] \\
&= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[\Big[\sum_{k=1}^p (\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^r \Big]^{\frac{1}{2r}} \Big] \\
&\leq \frac{1}{m} \Big[\mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[\sum_{k=1}^p (\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^r \Big] \Big]^{\frac{1}{2r}} \text{ (Jensen's inequality)} \\
&= \frac{1}{m} \Big[\sum_{k=1}^p \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[(\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^r \Big] \Big]^{\frac{1}{2r}}.
\end{aligned}$$

Assume that $r \geq 1$ is an integer, then, by Lemma 1, for any $k \in [1, p]$, we have

$$\mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[(\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^r \Big] \leq \Big(\eta_0 r \operatorname{Tr}[\mathbf{K}_k] \Big)^r.$$

Using these inequalities gives

$$\widehat{\mathfrak{R}}_S(H_p^1) \leq \frac{1}{m} \Big[\sum_{k=1}^p \Big(\eta_0 r \operatorname{Tr}[\mathbf{K}_k] \Big)^r \Big]^{\frac{1}{2r}} = \frac{\sqrt{\eta_0 r \|\mathbf{u}\|_r}}{m},$$

and concludes the proof. □

Theorem 2. *Let $p>1$ and assume that $K_k(x,x)\leq R^2$ for all $x\in\mathcal{X}$ and $k\in[1,p]$. Then, for any sample S of size m, the Rademacher complexity of the hypothesis set H_p^1 can be bounded as follows:*

$$\widehat{\mathfrak{R}}_S(H_p^1)\leq\sqrt{\frac{\eta_0 e\lceil\log p\rceil R^2}{m}}.$$

Proof. Since $K_k(x,x)\leq R^2$ for all $x\in\mathcal{X}$ and $k\in[1,p]$, $\mathrm{Tr}[\mathbf{K}_k]\leq mR^2$ for all $k\in[1,p]$. Thus, by Theorem 1, for any integer $r>1$, the Rademacher complexity can be bounded as follows

$$\widehat{\mathfrak{R}}_S(H_p^1)\leq\frac{1}{m}\left[p\left(\eta_0 rmR^2\right)^r\right]^{\frac{1}{2r}}=\sqrt{\frac{\eta_0 rp^{\frac{1}{r}}R^2}{m}}.$$

For $p>1$, the function $r\mapsto p^{1/r}r$ reaches its minimum at $r_0=\log p$, which gives $\widehat{\mathfrak{R}}_S(H_p^1)\leq\sqrt{\frac{\eta_0 e\lceil\log p\rceil R^2}{m}}$. □

Note that more generally, without assuming $K_k(x,x)\leq R^2$ for all k and all x, the same proof yields the following result:

$$\widehat{\mathfrak{R}}_S(H_p^1)\leq\sqrt{\frac{\eta_0 e\lceil\log p\rceil\|\mathbf{u}\|_\infty}{m}}.$$

Remarkably, the bound of the theorem has a very mild dependence on p. The theorem can be used to derive generalization bounds for learning kernels in classification, regression, and other tasks. We briefly illustrate its application to binary classification where the labels y are in $\{-1,+1\}$. Let $R(h)$ denote the generalization error of $h\in H_p^1$, that is $R(h)=\Pr[yh(x)<0]$. For a training sample $S=((x_1,y_1),\ldots,(x_m,y_m))$ and any $\rho>0$, define the ρ-empirical margin loss $\widehat{R}_\rho(h)$ as follows:

$$\widehat{R}_\rho(h)=\frac{1}{m}\sum_{i=1}^m\min\left(1,[1-y_ih(x_i)/\rho]_+\right).$$

Note that $\widehat{R}_\rho(h)$ is always upper bounded by the fraction of the training points with margin less than ρ:

$$\widehat{R}_\rho(h)\leq\frac{1}{m}\sum_{i=1}^m 1_{y_ih(x_i)<\rho}.$$

The following gives a margin-based generalization bound for the hypothesis set H_p^1.

Corollary 1. *Fix $\rho>0$. Then, for any integer $r>1$, for any $\delta>0$, with probability at least $1-\delta$, for any $h\in H_p^1$,*

$$R(h)\leq\widehat{R}_\rho(h)+\frac{2\sqrt{\eta_0 r}\|\mathbf{u}\|_r}{m\rho}+3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

with $\mathbf{u}=(\mathrm{Tr}[\mathbf{K}_1],\ldots,\mathrm{Tr}[\mathbf{K}_p])^\top$ and $\eta_0=\frac{23}{22}$.

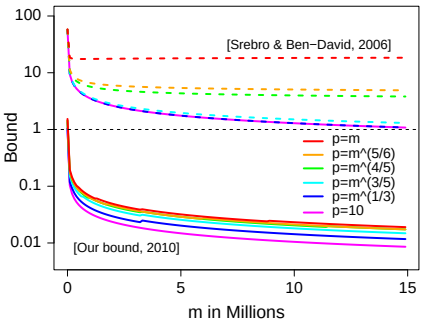


Figure 1. Plots of the bound of Srebro & Ben-David (2006) (dashed lines) and our new bounds (solid lines) as a function of the sample size m for $\delta=.01$ and $\rho/R=.2$. For these values and $m\leq 15\times 10^6$, the bound of Srebro and Ben-David is always above 1, it is of course converging for sufficiently large m. The plots for $p=10$ and $p=m^{1/3}$ roughly coincide in the case of the bound of Srebro & Ben-David (2006), which makes the first one not visible.

If additionally, $K_k(x,x)\leq R^2$ for all $x\in\mathcal{X}$ and $k\in[1,p]$, then, for $p>1$,

$$R(h)\leq\widehat{R}_\rho(h)+2\sqrt{\frac{\eta_0 e\lceil\log p\rceil R^2/\rho^2}{m}}+3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

Proof. With our definition of the Rademacher complexity, for any $\delta>0$, with probability at least $1-\delta$, the following bound holds for any $h\in H_p^1$ (Koltchinskii & Panchenko, 2002; Bartlett & Mendelson, 2002):

$$R(h)\leq\widehat{R}_\rho(h)+\frac{2}{\rho}\widehat{\mathfrak{R}}_S(H_p^1)+3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

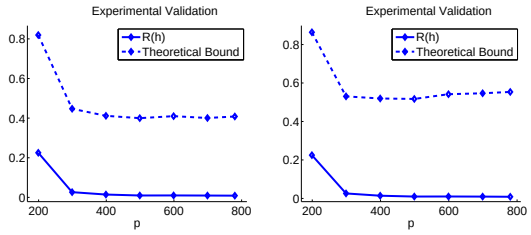
Plugging in the bound on the empirical Rademacher complexity given by Theorem 1 and Theorem 2 yields the statement of the corollary. □

The bound of the Corollary can be straightforwardly extended to hold uniformly over all choices of ρ, using standard techniques introduced by Koltchinskii & Panchenko (2002), at the price of the additional term $\frac{\log\log_2(4R/\rho)}{m}$ on the right-hand side.

The corollary gives a generalization bound for learning kernels with H_p^1 that is in

$$O\left(\sqrt{\frac{(\log p)\ R^2/\rho^2}{m}}\right).$$

In comparison, the best previous bound for learning kernels with convex combinations given by Srebro & Ben-David (2006) derived using the pseudo-dimension has a stronger



(a) L_1 Bound (b) L_2 Bound

Figure 2. Variation of the empirical test error and $R(h)$ as a function of the number of kernels, for $R(h)$ given by (a) Corollary 1 for L_1 regularization; (b) Corollary 2 for L_2 regularization. For these experiments, $m=36{,}000$, $\rho/R=.2$, and $\delta=.01$.

dependency with respect to p and is more complex:

$$O\left(\sqrt{8\frac{2+p\log\frac{128em^3R^2}{\rho^2 p}+256\frac{R^2}{\rho^2}\log\frac{\rho em}{8R}\log\frac{128mR^2}{\rho^2}}{m}}\right).$$

This bound is also not informative for $p > m$. Figure 1 compares the bound on $R(h)-\widehat{R}_\rho(h)$ obtained using this expression by Srebro and Ben-David with the new bound of Corollary 1, as a function of the sample size m. The comparison is made for different values of p, a normalized margin of $\rho/R = .2$ and the confidence parameter set to $\delta = .01$. Plots for different values of ρ/R are quite similar. As shown by the figure, larger values of p can significantly affect the bound of Srebro and Ben-David leading to quasi-flat plots for $p > m^{4/5}$. In comparison, the plots for our new bound show only a mild variation with p even for relatively large values such as $p \sim m$. Note also that, while the bound of Srebro and Ben-David does converge and becomes informative, its values, even for $p = 10$, are still above 1 for fairly large values of m. The new bound, in contrast, strongly encourages considering large numbers of base kernels in learning kernels. It was brought to our attention by an ICML reviewer that a bound similar to that of Theorem 2, with somewhat less favorable constants and for the expected value, was recently derived by Kakade et al. (2010) using a strong-convexity/smoothness argument.

Lower bound The $\sqrt{\log p}$ dependency of our generalization bound with respect to p cannot be improved upon. This can be seen by arguments in connection with the VC dimension lower bounds. Consider the case where the input space is $\mathcal{X}=\{\text{-}1,+1\}^p$ and where the feature mapping of each base kernel K_k, $k\in[1,p]$, is simply the canonical projection $\mathbf{x}\mapsto+x_k$ or $\mathbf{x}\mapsto-x_k$, where x_k is the kth component of $\mathbf{x}\in\mathcal{X}$. Thus, H_1^p then contains the hypothesis set $J^p=\{\mathbf{x}\mapsto sx_k\colon k\in[1,p], s\in\{\text{-}1,+1\}\}$ whose VC dimension is in $\Omega(\log p)$. For $\rho=1$ and $h\in J^p$, for any $x_i\in\mathcal{X}$, $y_ih(x_i)<\rho$ is equivalent to $y_ih(x_i)<0$. Thus, the empirical margin loss $\widehat{R}_\rho(h)$ coincides with the standard empirical error $\widehat{R}(h)$ for $h\in J^p$ and a margin bound with $\rho=1$ implies a standard generalization bound with the same complexity term. By the classical VC dimension lower bounds (Devroye & Lugosi, 1995; Anthony & Bartlett, 1999), that complexity term must be at least in $\Omega\big(\sqrt{\mathrm{VCDim}(J^p)/m}\big)=\Omega(\sqrt{\log p/m})$. A related simple example showing this lower bound was also suggested to us by N. Srebro.

We have also tested experimentally the behavior of the test error as a function of p and compared it to that of the theoretical bound given by Corollary 1 by learning with a large number of kernels $p\in[200,800]$, a sample size of $m=36{,}000$, and a normalized margin of $\rho/R=.2$. These results are for rank-1 base kernels generated from individual features of the MNIST dataset (http://yann.lecun.com/exdb/mnist/). The magnitude of each kernel weight is chosen proportionally to the correlation of the corresponding feature with the training labels. The results show that the behavior of the test error as a function of p matches the one predicted by our bound, see Figure 2(a).

4. Rademacher complexity bound for H_p^q

This section presents bounds on the Rademacher complexity of the hypothesis sets H_p^q for various values of $q>1$, including $q=2$.

Theorem 3. *Let $q,r\geq1$ with $\frac{1}{q}+\frac{1}{r}=1$ and assume that r is an integer. Then, for any sample S of size m, the empirical Rademacher complexity of the hypothesis set H_p^q can be bounded as follows:*

$$\widehat{\mathfrak{R}}_S(H_p^q)\leq\frac{\sqrt{\eta_0 r\|\mathbf{u}\|_r}}{m},$$

where $\mathbf{u}=(\mathrm{Tr}[\mathbf{K}_1],\ldots,\mathrm{Tr}[\mathbf{K}_p])^\top$ and $\eta_0=\frac{23}{22}$.

Proof. By Proposition 1, $\widehat{\mathfrak{R}}_S(H_p^q)=\frac{1}{m}\,\mathrm{E}_{\boldsymbol{\sigma}}\left[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_r}\right]$. with $\mathbf{u}_{\boldsymbol{\sigma}}=(\boldsymbol{\sigma}^\top\mathbf{K}_1\boldsymbol{\sigma},\ldots,\boldsymbol{\sigma}^\top\mathbf{K}_p\boldsymbol{\sigma})^\top$. The rest of the proof is identical to that of Theorem 1: using Jensen's inequality and Lemma 1, which applies because r is an integer, we obtain similarly

$$\widehat{\mathfrak{R}}_S(H_p^q)\leq\frac{1}{m}\Big[\sum_{k=1}^p\Big(\eta_0 r\,\mathrm{Tr}[\mathbf{K}_k]\Big)^r\Big]^{\frac{1}{2r}}.$$ □

In particular, for $q=r=2$, the theorem implies

$$\widehat{\mathfrak{R}}_S(H_p^2)\leq\frac{\sqrt{2\eta_0\|\mathbf{u}\|_2}}{m}.$$

Theorem 4. *Let $q,r\geq1$ with $\frac{1}{q}+\frac{1}{r}=1$ and assume that r is an integer. Let $p>1$ and assume that $K_k(x,x)\leq R^2$ for all $x\in\mathcal{X}$ and $k\in[1,p]$. Then, for any sample S of size m, the Rademacher complexity of the hypothesis set H_p^q can*

be bounded as follows:

$$\widehat{\mathfrak{R}}_S(H_p^q) \leq \sqrt{\frac{\eta_0 r p^{\frac{1}{r}} R^2}{m}}.$$

Proof. Since $K_k(x,x) \leq R^2$ for all $x \in \mathcal{X}$ and $k \in [1,p]$, $\mathrm{Tr}[\mathbf{K}_k] \leq mR^2$ for all $k \in [1,p]$. Thus, by Theorem 3, the Rademacher complexity can be bounded as follows

$$\widehat{\mathfrak{R}}_S(H_p^q) \leq \frac{1}{m}\left[p\left(\eta_0 r m R^2\right)^r\right]^{\frac{1}{2r}} = \sqrt{\frac{\eta_0 r p^{\frac{1}{r}} R^2}{m}}. \qquad \square$$

The bound of the theorem has only a mild dependence ($\sqrt[2r]{\cdot}$) on the number of kernels p. In particular, for $q = r = 2$, under the assumptions of the theorem,

$$\widehat{\mathfrak{R}}_S(H_p^2) \leq \sqrt{\frac{2\eta_0 \sqrt{p} R^2}{m}},$$

and the dependence is in $O(p^{1/4})$.

Proceeding as in the L_1 case leads to the following margin bound in binary classification.

Corollary 2. *Let $q, r \geq 1$ with $\frac{1}{q}+\frac{1}{r}=1$ and assume that r is an integer. Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1-\delta$, for any $h \in H_p^q$,*

$$R(h) \leq \widehat{R}_\rho(h) + \frac{2\sqrt{\eta_0 r \|\mathbf{u}\|_r}}{m\rho} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

with $\mathbf{u} = (\mathrm{Tr}[\mathbf{K}_1], \ldots, \mathrm{Tr}[\mathbf{K}_p])^\top$ and $\eta_0 = \frac{23}{22}$.

If additionally, $K_k(x,x) \leq R^2$ for all $x \in \mathcal{X}$ and $k \in [1,p]$, then, for $p > 1$,

$$R(h) \leq \widehat{R}_\rho(h) + 2p^{\frac{1}{2r}}\sqrt{\frac{\eta_0 r R^2/\rho^2}{m}} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

In particular, for $q = r = 2$, the generalization bound of the corollary becomes

$$R(h) \leq \widehat{R}_\rho(h) + 2p^{\frac{1}{4}}\sqrt{\frac{2\eta_0 R^2/\rho^2}{m}} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Figure 3 shows a comparison of the L_2 regularization bound of this corollary with the L_1 regularization bound of Corollary 1. As can be seen from the plots, the two bounds are very close for smaller values of p. For larger values ($p \sim m$), the difference becomes significant. The bound for L_2 regularization is converging for these values but at a slower rate of $O\left(\frac{R/\rho}{m^{1/4}}\right)$.

As with the L_1 bound we also tested experimentally the behavior of the test error as a function of p and compared it to that of the theoretical bound given by Corollary 2 by learning with a large number of kernels. Again, our results show that the behavior of the test error as a function of p matches the one predicted by our bound, see Figure 2(b).

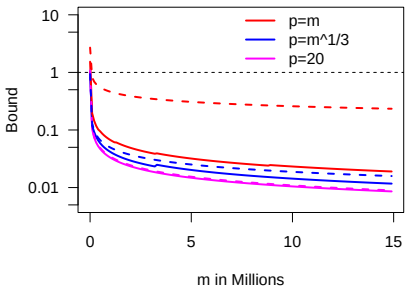


Figure 3. Comparison of the L_1 regularization bound of Corollary 1 and the L_2 regularization bound of Corollary 2 (dotted lines) as a function of the sample size m for $\delta = .01$ and $\rho/R = .2$. For $p = 20$, the L_1 and L_2 bounds roughly coincide.

Lower bound The $p^{1/(2r)}$ dependency of the generalization bound of Corollary 2 cannot be improved. In particular, the $p^{1/4}$ dependency is tight for the hypothesis set H_p^2. This is clear since in particular when all p kernel functions are equal, $\sum_{k=1}^p \mu_k K_k = (\sum_{k=1}^p \mu_k) K_1 \leq p^{1/r} K_1$. H_p^q then coincides with the set of functions in H_1^q each multiplied by $p^{1/(2r)}$.

5. Proof techniques

Our proof techniques are somewhat general and apply similarly to other problems. In particular, they can be used as alternative methods to derive bounds on the Rademacher complexity of linear functions classes, such as those given by Kakade et al. (2009), using strong convexity. In fact, in some cases, they can lead to similar bounds but with tighter constants. The following theorem illustrates that in the case of linear functions constrained by the norm $\|\cdot\|_q$.

Theorem 5. *Let $q, r \geq 1$ with $\frac{1}{q}+\frac{1}{r}=1$, r an even integer such that $r \geq 2$. Let $\mathcal{X} = \{x : \|x\|_r \leq X\}$, and let $\mathcal{F}$ be the class of linear functions over $\mathcal{X}$ defined by $\mathcal{F} = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} : \|\mathbf{w}\|_q \leq W\}$, then, for any sample $S = (x_1, \ldots, x_m)$, the following bound holds for the empirical Rademacher complexity of this class:*

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) \leq XW\sqrt{\frac{\eta_0 r}{2m}}.$$

Clearly, this immediately yields the same bound on the Rademacher complexity $\mathfrak{R}_m(\mathcal{F}) = \mathrm{E}_S[\widehat{\mathfrak{R}}_S(\mathcal{F})]$. The bound given by Kakade et al. (2009)[Section 3.1] in this case is $\mathfrak{R}_m(\mathcal{F}) \leq XW\sqrt{\frac{r-1}{m}}$. Since $\eta_0 r/2 \leq r-1$, for an even integer $r > 2$, our bound is always tighter.

Proof. The proof is similar to and uses that of Theorem 1. By the definition of the dual norms, the following holds:

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) = \frac{1}{m}\mathrm{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_q \leq W} \sum_{i=1}^m \sigma_i \mathbf{w} \cdot \mathbf{x}_i\right] = \frac{W}{m}\mathrm{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^m \sigma_i \mathbf{x}_i\right\|_r\right].$$

By Jensen's inequality,

$$\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big\|\sum_{i=1}^{m}\sigma_i\mathbf{x}_i\Big\|_r \leq \Big[\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big\|\sum_{i=1}^{m}\sigma_i\mathbf{x}_i\Big\|_r^r\Big]^{\frac{1}{r}} = \Big[\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\sum_{j=1}^{N}\Big[\sum_{i=1}^{m}\sigma_i x_{ij}\Big]^r\Big]^{\frac{1}{r}},$$

where we denote by N the dimension of the space and by x_{ij} the jth coordinate of $\mathbf{x}_i$. Now, we can bound the term $\mathrm{E}_{\boldsymbol{\sigma}}\left[\left[\sum_{i=1}^{m}\sigma_i x_{ij}\right]^r\right]$ using Lemma 1 and obtain:

$$\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\Big[\sum_{i=1}^{m}\sigma_i x_{ij}\Big]^r\Big] = \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\Big[\sum_{i,l=1}^{m}\sigma_i\sigma_l x_{ij}x_{lj}\Big]^{\frac{r}{2}}\Big] \leq \Big[\frac{\eta_0 r}{2}\sum_{i=1}^{m}x_{ij}^2\Big]^{\frac{r}{2}}.$$

Thus,

$$\begin{aligned}\widehat{\mathfrak{R}}_S(\mathcal{F}) &\leq \frac{W}{m}\Big(\frac{\eta_0 r}{2}\Big)^{1/2}\Big[\sum_{j=1}^{N}\Big(\sum_{i=1}^{m}x_{ij}^2\Big)^{r/2}\Big]^{\frac{1}{r}}\\ &= W\sqrt{\frac{\eta_0 r}{2m}}\Big[\sum_{j=1}^{N}\Big(\frac{1}{m}\sum_{i=1}^{m}x_{ij}^2\Big)^{r/2}\Big]^{\frac{1}{r}}.\end{aligned}$$

Since $r \geq 2$, by Jensen's inequality, $\left(\frac{1}{m}\sum_{i=1}^{m}x_{ij}^2\right)^{r/2} \leq \frac{1}{m}\sum_{i=1}^{m}x_{ij}^r$. Thus,

$$\begin{aligned}\widehat{\mathfrak{R}}_S(\mathcal{F}) &\leq W\sqrt{\frac{\eta_0 r}{2m}}\Big[\sum_{j=1}^{N}\frac{1}{m}\sum_{i=1}^{m}x_{ij}^r\Big]^{\frac{1}{r}}\\ &= W\sqrt{\frac{\eta_0 r}{2m}}\Big[\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}_i\|_r^r\Big]^{\frac{1}{r}} \leq W\sqrt{\frac{\eta_0 r}{2m}}X. \quad \square\end{aligned}$$

6. Conclusion

We presented several new generalization bounds for the problem of learning kernels with non-negative combinations of base kernels and outlined the relevance of our proof techniques to the analysis of the complexity of the class of linear functions. Our bounds are simpler and significantly improve over previous bounds. Their behavior matches empirical observations with a large number of base kernels. Their very mild dependency on the number of kernels suggests the use of a large number of kernels for this problem. Recent experiments by Cortes et al. (2009a; 2010) in regression using a large number of kernels seems to corroborate this idea. Much needs to be done however to combine these theoretical findings with the somewhat disappointing performance observed in practice in most learning kernel experiments.

Acknowledgments

We thank ICML reviewers for insightful comments on an earlier draft of this paper and N. Srebro for discussions.

References

Anthony, Martin and Bartlett, Peter L. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Argyriou, Andreas, Micchelli, Charles, and Pontil, Massimiliano. Learning convex combinations of continuously parameterized basic kernels. In *COLT*, 2005.

Argyriou, Andreas, Hauser, Raphael, Micchelli, Charles, and Pontil, Massimiliano. A DC-programming algorithm for kernel selection. In *ICML*, 2006.

Bach, F. Exploring large feature spaces with hierarchical multiple kernel learning. In *NIPS 2009*, 2008.

Bartlett, Peter L. and Mendelson, Shahar. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:2002, 2002.

Boser, Bernhard, Guyon, Isabelle, and Vapnik, Vladimir. A training algorithm for optimal margin classifiers. In *COLT*, 1992.

Bousquet, Olivier and Herrmann, Daniel J. L. On the complexity of learning the kernel matrix. In *NIPS*, 2002.

Cortes, Corinna and Vapnik, Vladimir. Support-Vector Networks. *Machine Learning*, 20(3), 1995.

Cortes, Corinna, Mohri, Mehryar, and Rostamizadeh, Afshin. L_2 regularization for learning kernels. In *UAI 2009*, 2009a.

Cortes, Corinna, Mohri, Mehryar, and Rostamizadeh, Afshin. Learning non-linear combinations of kernels. In *NIPS 2009*. MIT Press, 2009b.

Cortes, Corinna, Mohri, Mehryar, and Rostamizadeh, Afshin. Two-stage learning kernel methods. In *ICML 2010*, 2010.

Devroye, Luc and Lugosi, Gábor. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28(7), 1995.

Kakade, Sham M., Sridharan, Karthik, and Tewari, Ambuj. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, 2009.

Kakade, Sham M., Shalev-Shwartz, Shai, and Tewari, Ambuj. Applications of strong convexity–strong smoothness duality to learning with matrices, 2010. arXiv:0910.0610v1.

Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.

Koltchinskii, Vladimir and Yuan, Ming. Sparse recovery in large ensembles of kernel machines on-line learning and bandits. In *COLT*, 2008.

Lanckriet, Gert, Cristianini, Nello, Bartlett, Peter, Ghaoui, Laurent El, and Jordan, Michael. Learning the kernel matrix with semidefinite programming. *JMLR*, 5, 2004.

Lewis, Darrin P., Jebara, Tony, and Noble, William Stafford. Nonstationary kernel combination. In *ICML*, 2006.

Micchelli, Charles and Pontil, Massimiliano. Learning the kernel function via regularization. *JMLR*, 6, 2005.

Ong, Cheng Soon, Smola, Alex, and Williamson, Robert. Learning the kernel with hyperkernels. *JMLR*, 6, 2005.

Schölkopf, Bernhard and Smola, Alex. *Learning with Kernels*. MIT Press: Cambridge, MA, 2002.

Shawe-Taylor, John and Cristianini, Nello. *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, 2004.

Srebro, Nathan and Ben-David, Shai. Learning bounds for support vector machines with learned kernels. In *COLT*, 2006.

Vapnik, Vladimir N. *Statistical Learning Theory*. John Wiley & Sons, 1998.

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

Zien, Alexander and Ong, Cheng Soon. Multiclass multiple kernel learning. In *ICML 2007*, 2007.

A. Bound on Multinomial Coefficients

In the proof of Lemma 1, we need to upper bound the ratio $\binom{2r'}{2t_1,\ldots,2t_m}/\binom{r'}{t_1,\ldots,t_m}$. The following rough but straightforward inequality is sufficient to derive a bound on the Rademacher complexity with somewhat less favorable constants:

$$\begin{aligned}\binom{2r'}{2t_1,\ldots,2t_m} &= \frac{(2r')!}{(2t_1)!\cdots(2t_m)!} \leq \frac{(2r')!}{(t_1)!\cdots(t_m)!} \\ &\leq \frac{(2r')^{r'}\cdot r'!}{(t_1)!\cdots(t_m)!} = (2r')^{r'}\binom{r'}{t_1,\ldots,t_m}.\end{aligned}$$

To further improve this result, the next lemma uses Stirling's approximation valid for all $n \geq 1$: $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\lambda_n}$, with $\frac{1}{12n+1} < \lambda_n < \frac{1}{12n}$.

Lemma 2. *For all $r' > 0$ and $t_1, \ldots, t_m$, it holds that:*

$$\binom{2r'}{2t_1,\ldots,2t_m} \leq \left((1+\tfrac{1}{22})r'\right)^{r'}\binom{r'}{t_1,\ldots,t_m}.$$

Proof. By Stirling's formula,

$$\begin{aligned}\frac{(2r')!}{r'!} &= \sqrt{2}\left(\frac{2r'}{e}\right)^{2r'}\left(\frac{r'}{e}\right)^{-r'} e^{\lambda_{2r'}-\lambda_{r'}} \qquad (3)\\ &= \sqrt{2}\,2^{2r'}\left(\frac{r'}{e}\right)^{r'} e^{\lambda_{2r'}-\lambda_{r'}} = \sqrt{2}\left(\frac{4r'}{e}\right)^{r'} e^{\lambda_{2r'}-\lambda_{r'}}.\end{aligned}$$

Similarly, for any $t_i \geq 1$, we can write

$$\frac{t_i!}{(2t_i)!} = \frac{1}{\sqrt{2}}\left(\frac{e}{4t_i}\right)^{t_i} e^{\lambda_{t_i}-\lambda_{2t_i}} \leq \frac{1}{\sqrt{2}}\left(\frac{e}{4}\right)^{t_i} e^{\lambda_{t_i}-\lambda_{2t_i}}.$$

Using $\sum_{i=1}^m t_i = \sum_{t_i\geq 1} t_i = r'$, we obtain:

$$\prod_{t_i\geq 1}\frac{t_i!}{(2t_i)!} \leq \frac{1}{\sqrt{2}}\left(\frac{e}{4}\right)^{r'} e^{\sum_{t_i\geq 1}(\lambda_{t_i}-\lambda_{2t_i})}. \qquad (4)$$

In view of Eqn 3 and 4, the following inequality holds:

$$\binom{2r'}{2t_1,\ldots,2t_m}/\binom{r'}{t_1,\ldots,t_m} \leq (r')^{r'} e^{\lambda_{2r'}-\lambda_{r'}+\sum_{t_i\geq 1}(\lambda_{t_i}-\lambda_{2t_i})}.$$

We now derive an upper bound on the terms appearing in the exponent. Using the inequalities imposed on λ_{t_i} and λ_{2t_i} and the fact that the sum of t_is is r' leads to:

$$\begin{aligned}\sum_{t_i\geq 1}\lambda_{t_i}-\lambda_{2t_i} &\leq \sum_{t_i\geq 1}\frac{1}{12t_i}-\frac{1}{24t_i+1} = \sum_{t_i\geq 1}\frac{12t_i+1}{12t_i[24t_i+1]} \\ &\leq \sum_{t_i\geq 1}\frac{1+\frac{1}{12}}{24t_i+1} \leq \sum_{t_i\geq 1}\frac{\frac{13}{12}}{25} \leq \frac{13r'}{300},\end{aligned}$$

and $\lambda_{2r'}-\lambda_{r'} \leq \frac{1}{24r'}-\frac{1}{12r'+1} \leq 0$. The inequality $e^{13/300} < 1 + 1/22$ then yields the statement of the lemma. □

B. Proof of Lemma 1

Proof. Since r is an integer, we can write:

$$\begin{aligned}&\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[(\boldsymbol{\sigma}^\top\mathbf{K}\boldsymbol{\sigma})^r\right] = \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\Big(\sum_{i,j=1}^m \sigma_i\sigma_j K_k(x_i,x_j)\Big)^r\right] \\ &= \sum_{\substack{1\leq i_1,\ldots,i_r\leq m\\ 1\leq j_1,\ldots,j_r\leq m}} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\prod_{s=1}^r \sigma_{i_s}\sigma_{j_s}\right]\prod_{s=1}^r K_k(x_{i_s},x_{j_s}) \\ &\leq \sum_{\substack{1\leq i_1,\ldots,i_r\leq m\\ 1\leq j_1,\ldots,j_r\leq m}} \left|\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\prod_{s=1}^r \sigma_{i_s}\sigma_{j_s}\right]\right|\prod_{s=1}^r |K_k(x_{i_s},x_{j_s})| \\ &\leq \sum_{\substack{1\leq i_1,\ldots,i_r\leq m\\ 1\leq j_1,\ldots,j_r\leq m}} \left|\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\prod_{s=1}^r \sigma_{i_s}\sigma_{j_s}\right]\right| \\ &\qquad \prod_{s=1}^r \sqrt{K_k(x_{i_s},x_{i_s})K_k(x_{j_s},x_{j_s})} \quad \text{(Cauchy-Schwarz)} \\ &= \sum_{s_1+\ldots+s_m=2r}\binom{2r}{s_1,\ldots,s_m}\left|\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}[\sigma_1^{s_1}\cdots\sigma_m^{s_m}]\right|\prod_{i=1}^m\sqrt{K_k(x_i,x_i)^{s_i}}.\end{aligned}$$

Since $\mathrm{E}[\sigma_i] = 0$ for all i and since the Rademacher variables are independent, we can write $\mathrm{E}[\sigma_{i_1}\ldots\sigma_{i_l}] = \mathrm{E}[\sigma_{i_1}]\cdots\mathrm{E}[\sigma_{i_l}] = 0$ for any l distinct variables $\sigma_{i_1},\ldots,\sigma_{i_l}$. Thus, $\mathrm{E}_{\boldsymbol{\sigma}}\left[\sigma_1^{s_1}\cdots\sigma_1^{s_m}\right] = 0$ unless all s_is are even, in which case $\mathrm{E}_{\boldsymbol{\sigma}}\left[\sigma_1^{s_1}\cdots\sigma_m^{s_m}\right] = 1$. It follows that:

$$\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[(\boldsymbol{\sigma}^\top\mathbf{K}\boldsymbol{\sigma})^r\right] \leq \sum_{2t_1+\ldots+2t_m=2r}\binom{2r}{2t_1,\ldots,2t_m}\prod_{i=1}^m K_k(x_i,x_i)^{t_i}.$$

By Lemma 2, each multinomial coefficient $\binom{2r}{2t_1,\ldots,2t_m}$ can be bounded by $(\eta_0 r)^r\binom{r}{t_1,\ldots,t_m}$, where $\eta_0 = \frac{23}{22}$. This gives

$$\begin{aligned}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[(\boldsymbol{\sigma}^\top\mathbf{K}\boldsymbol{\sigma})^r\right] &\leq (\eta_0 r)^r\sum_{t_1+\ldots+t_m=r}\binom{r}{t_1,\ldots,t_m}\prod_{i=1}^m K_k(x_i,x_i)^{t_i} \\ &= (\eta_0 r)^r(\mathrm{Tr}[\mathbf{K}])^r = \Big(\eta_0 r\,\mathrm{Tr}[\mathbf{K}]\Big)^r,\end{aligned}$$

which is the statement of the lemma. □