Improved Loss Bounds For Multiple Kernel Learning

Zakria Hussain & John Shawe-Taylor
Centre for Computational Statistics and Machine Learning
Department of Computer Science
University College London
{z.hussain, jst}@cs.ucl.ac.uk

Abstract

We propose two new generalization error bounds for multiple kernel learning (MKL). First, using the bound of Srebro and Ben-David (2006) as a starting point, we derive a new version which uses a simple counting argument for the choice of kernels in order to generate a tighter bound when 1-norm regularization (sparsity) is imposed in the kernel learning problem. The second bound is a Rademacher complexity bound which is *additive* in the (logarithmic) kernel complexity and margin term. This dependence is superior to all previously published Rademacher bounds for learning a convex combination of kernels, including the recent bound of Cortes et al. (2010), which exhibits a multiplicative interaction. We illustrate the tightness of our bounds with simulations.

1 INTRODUCTION

Bounds for multiple kernel learning (MKL) have proved a popular research direction since the introduction of MKL algorithms (Lanckriet et al., 2004; Bach et al., 2004; Argyriou et al., 2005). The first bound was in the seminal MKL paper of Lanckriet et al. (2004), and was a Rademacher complexity bound (Koltchinskii and Panchenko, 2002; Bartlett and Mendelson, 2002). Bounds for MKL were also proposed by Micchelli et al. (2005) and Ying and Zhou (2007). Later, Srebro and Ben-David (2006) improved on these bounds, and showed the bound of Lanckriet et al. (2004) to be vacuous in the case of a linear or convex combination of kernels (not including the 1-norm case). The bound of Srebro and Ben-David (2006) was important due to the fact that it scaled additively in the kernel complexity term p and the margin term γ. It is believed that this scaling factor should be the defacto standard for MKL algorithms, as any multiplicative interaction between these two terms would blow up considerably, requiring a multiplicative increase in the sample size and resulting in a loose bound. Recently, Cortes et al. (2010) have proposed an MKL bound for a family of convex combination of base kernels using Rademacher complexity, which scales multiplicatively between these two terms. However, the dependence on the kernel complexity term is only logarithmic. Hence, when learning a convex combination of kernels, their bound can be considerably tighter than the Srebro and Ben-David (2006) result.

In this paper we present two new results. The first result shows that we can obtain a logarithmic dependency of the kernel complexity p for the Srebro and Ben-David (2006) bound in the case of a (sparse) linear/convex combination of kernels. However, at the cost of an additional integer $d \ll p$ corresponding to the number of chosen kernels in the final solution. However, d is typically much smaller than p and so our bound can be tighter than the bound of Srebro and Ben-David (2006) when an MKL algorithm chooses a sparse number of kernels from a large finite family of kernels *i.e.*, ℓ_1 block-norm regularization of MKL. Our second result is a Rademacher complexity bound which improves on the bound of Cortes et al. (2010), in that we now have an additive dependency between the logarithmic kernel complexity term $\ln p$ and the margin term γ. To our knowledge, when choosing a convex combination of kernels using MKL, this result is tighter than any previously published bounds.

The rest of the paper is structured as follows. In the following section we formalize our contributions. After discussing preliminary definitions in Section 3, we move onto the main proofs of the paper given in Section 4 and Section 5, discussing the sparsity and

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

Rademacher bounds respectively. We conclude in Section 6.

2 OUR CONTRIBUTION

Let $p \in \mathbb{N}$ denote the complexity of a family of kernels, $\gamma \in [0,1]$ the margin and $m \in \mathbb{N}$ the size of a sample. Srebro and Ben-David (2006) have proposed a multiple kernel learning bound whose estimation error can be upper bounded by a term of the order $\tilde{\mathcal{O}}\left(\sqrt{(p+1/\gamma^2)/m}\right)$, where $\tilde{\mathcal{O}}$ hides logarithmic factors (see Section 4 for the exact form of the bound). Typically in MKL (Lanckriet et al., 2004), we find a convex/linear combination of $d \ll p$ kernels in the final solution (sparsity using ℓ_1 norm regularization). Therefore, taking the result of Srebro and Ben-David (2006) as a starting point we apply a counting argument over the choice of kernels plus a union bound in order to construct a bound of the order $\tilde{\mathcal{O}}\left(\sqrt{(\log p + 1/\gamma^2 + 2d)/m}\right)$. This gives us a logarithmic dependency in the kernel complexity term p and is tighter when $\log p + 2d < p$.

Cortes et al. (2010) have recently proposed a bound using Rademacher complexity of the order $\mathcal{O}\left(\sqrt{((\ln p)1/\gamma^2)/m}\right)$, showing it to be tighter than the bound of Srebro and Ben-David (2006), when learning a 1-norm regularized convex combination of kernels. We propose a Rademacher complexity bound (with a simpler proof than that of Cortes et al. (2010)) which is of the order $\mathcal{O}\left(\sqrt{(\ln p + 1/\gamma^2)/m}\right)$. As far as we are aware this is the first Rademacher complexity bound for MKL which is *additive* in the kernel complexity term p and margin term γ. This is tighter than the bounds of Srebro and Ben-David (2006) and Cortes et al. (2010) for learning kernels of convex combinations, both of which are considered the current state-of-the-art.

3 PRELIMINARIES

Let $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ be an m-sample where $x_i \in \mathcal{X} \subset \mathbb{R}^n$ and $y_i \in \mathcal{Y} = \{-1, +1\}$, with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\mathbf{x} = \{x_1, \ldots, x_m\}$ contain the input vectors.

Definition 1 (Aizerman et al., 1964)**.** *A kernel is a function κ that for all $x, x' \in \mathcal{X}$ satisfies*

$$\kappa(x, x') = \langle \phi(x), \phi(x') \rangle,$$

where ϕ is a mapping from $\mathcal{X}$ to an (inner product) Hilbert space $\mathcal{H}$

$$\phi : \mathcal{X} \mapsto \mathcal{H}.$$

Kernel learning algorithms (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004) make use of the $m \times m$ kernel matrix $K = [\kappa(x_i, x_j)]_{i,j=1}^m$ defined using the training inputs $\mathbf{x}$. When using the kernel representation it is not always possible to represent the weight vector w explicitly and so we can use the function f directly as the predictor:

$$f(x) = \sum_{i=1}^m \alpha_i \kappa(x_i, x),$$

where $\alpha = (\alpha_1, \ldots, \alpha_m)$ is the dual weight vector. Given a kernel κ, learning can be described as finding a function from the class of functions (Srebro and Ben-David, 2006):

$$\mathcal{F} = \{x \mapsto \langle w, \phi(x) \rangle \mid \|w\|_2 \leq 1, \kappa(x, x') = \langle \phi(x), \phi(x') \rangle\}$$

minimizing the hinge loss

$$\hat{h}^\gamma(f) = \frac{1}{m} \sum_{i=1}^m \xi_i,$$

where $\xi_i = \max(\gamma - y_i f(x_i), 0)$. We call $\gamma \in [0,1]$ the *margin*.

For the generalization error bounds we assume that the data are generated iid from a fixed but unknown probability distribution $\mathcal{D}$ over the joint space $\mathcal{X} \times \mathcal{Y}$. Given the *true error* of a function f:

$$\mathrm{err}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}(y f(x) \leq 0),$$

the *empirical margin error* of f with margin $\gamma > 0$:

$$\hat{\mathrm{err}}^\gamma(f) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(y_i f(x_i) < \gamma)$$

where $\mathbb{I}$ is the indicator function, and the estimation error $\mathrm{est}^\gamma(f)$

$$\mathrm{est}^\gamma(f) = |\mathrm{err}(f) - \hat{\mathrm{err}}^\gamma(f)|,$$

we would like to find an upper bound for $\mathrm{est}^\gamma(f)$. In the sequel we will state the bounds in standard form, where the true error $\mathrm{err}(f)$ of a function f is upper bounded by the empirical margin error $\hat{\mathrm{err}}^\gamma(f)$ plus the estimation error $\mathrm{est}^\gamma(f)$:

$$\mathrm{err}(f) \leq \hat{\mathrm{err}}^\gamma(f) + \mathrm{est}^\gamma(f). \tag{1}$$

Let $\mathcal{K} = \{\kappa_1, \ldots, \kappa_p\}$ denote a family of kernels, where each kernel κ_i is called the ith *base* kernel. The following kernel families are formed using a linear or convex combination of base kernels:

$$\mathcal{K}_{\mathrm{lin}}(\kappa_1, \ldots, \kappa_p) = \left\{ \kappa^{\boldsymbol{\eta}} = \sum_{i=1}^p \eta_i \kappa_i \mid K^{\boldsymbol{\eta}} \succcurlyeq 0, \sum_{i=1}^p \eta_i = 1 \right\}$$

$$\mathcal{K}_{\mathrm{con}}(\kappa_1, \ldots, \kappa_p) = \left\{ \kappa^{\boldsymbol{\eta}} = \sum_{i=1}^p \eta_i \kappa_i \mid \eta_i \geq 0, \sum_{i=1}^p \eta_i = 1 \right\}.$$

These two kernel families are considered *finite* dimensional – hence p is the complexity of the kernel family (*i.e.*, cardinality of the set). The MKL problem can be described as finding a function f from the class:

$$\mathcal{F}_{\mathcal{K}} = \cup_{\kappa_j \in \mathcal{K}} \mathcal{F}_j,$$

that minimizes $\hat{h}^{\gamma}(f)$, where $j \in \{1, \ldots, p\}$.

4 IMPROVED MARGIN BOUND FOR SPARSE MKL

In this section we use covering number bounds for learning the kernel with support vector machines (SVM) (Srebro and Ben-David, 2006). We derive an upper bound for the above finite dimensional kernel families (that use ℓ_1 norm regularization *i.e.*, sparsity) using covering numbers, potentially resulting in tighter bounds than Srebro and Ben-David (2006), when a sparse number of kernels is present in the final combination (*i.e.*, Lanckriet et al. (2004); Bach (2009)). Before presenting our first result we define covering numbers for kernels and the bound of Srebro and Ben-David (2006).

Definition 2 (covering number)**.** *A subset $\tilde{A} \subset A$ is an ϵ-cover of A under the metric $d(\cdot, \cdot)$ if for any $a \in A$ there exists $\tilde{a} \in \tilde{A}$ with $d(a, \tilde{a}) \leq \epsilon$. The* covering number $\mathcal{N}_d(A, \epsilon)$ *is the size of the smallest ϵ-cover of A.*

Given a sample of m inputs $\mathbf{x}$ we can define the following ℓ_∞ metric:

$$d_\infty^{\mathbf{x}}(f_1, f_2) = \max_{1 \leq i \leq m} |f_1(x_i) - f_2(x_i)|.$$

The uniform ℓ_∞ covering number $\mathcal{N}_m(\mathcal{F}, \epsilon)$ of a predictor class $\mathcal{F}$ is given by considering all possible inputs $\mathbf{x}$ of size m:

$$\mathcal{N}_m(\mathcal{F}, \epsilon) = \sup_{|\mathbf{x}|=m} \mathcal{N}_{d_\infty^{\mathbf{x}}}(\mathcal{F}, \epsilon).$$

In the kernel learning scenario we have:

$$D_\infty^{\mathbf{x}}(\kappa, \tilde{\kappa}) = \max_{1 \leq i,j \leq m} |\kappa(x_i, x_j) - \tilde{\kappa}(x_i, x_j)|.$$

Theorem 1 (Srebro and Ben-David, 2006)**.** *Fix $\gamma > 0$ and $\delta \in (0, 1)$. For any kernel family $\mathcal{K}$, bounded by $R^2 \geq \kappa(x, x)$ and with pseudo-dimension p, with probability at least $1 - \delta$ over the choice of a random training set $\mathbf{z} = \{z_i\}_{i=1}^m$ of size m we have, for any $f \in \mathcal{F}_{\mathcal{K}}$:*

$$\mathrm{err}(f) \leq \hat{\mathrm{err}}^{\gamma}(f) + \mathrm{est}^{\gamma}(f),$$

where

$$\mathrm{est}^{\gamma}(f) = \sqrt{8 \frac{2 + p \log \frac{128 e m^3 R^2}{\gamma^2 p} + 256 \frac{R^2}{\gamma^2} \log \frac{\gamma e m}{8R} \log \frac{128 m R^2}{\gamma^2} - \log \delta}{m}}.$$

We can apply a counting argument to this bound if our algorithm chooses $d < p$ kernels from a family of finite kernels $\mathcal{K} = \{\kappa_1, \ldots, \kappa_p\}$, together with a union bound over all choices of kernels. Therefore, the following result is applicable to the kernel families $\mathcal{K}_{\mathrm{lin}}$ and $\mathcal{K}_{\mathrm{con}}$, described in Section 3.

Theorem 2. *Fix $\gamma > 0$ and $\delta \in (0, 1)$. Let $d < p$ be the number of kernels involved in the final MKL solution. For any finite kernel family $\mathcal{K} = \{\kappa_1, \ldots, \kappa_p\}$, bounded by R^2, with probability at least $1 - \delta$ over the choice of a random training set $\mathbf{z} = \{z_i\}_{i=1}^m$ of size m we have, for any $f \in \mathcal{F}_{\mathcal{K}}$:*

$$\mathrm{err}(f) \leq \hat{\mathrm{err}}^{\gamma}(f) + \mathrm{est}^{\gamma}(f),$$

where

$$\mathrm{est}^{\gamma}(f) = \sqrt{8 \frac{2 + d \log \frac{ep}{d} + d \log \frac{128 e m^3 R^2}{\gamma^2 d} + 256 \frac{R^2}{\gamma^2} \log \frac{\gamma e m}{8R} \log \frac{128 m R^2}{\gamma^2} - \log \frac{\delta}{p}}{m}}.$$

Proof. From Anthony and Bartlett (1999) (Theorem 10.1) we have:

$$\sup_{f \in \mathcal{F}} \mathrm{est}^{\gamma}(f) \leq \sqrt{8 \frac{1 + \log \mathcal{N}_{2m}(\mathcal{F}, \gamma/2) - \log \delta}{m}},$$

which is found by solving the following equation for $\epsilon > 0$:

$$2\mathcal{N}_{2m}(\mathcal{F}, \gamma/2) \exp\left(-\frac{\epsilon^2 m}{8}\right) = \delta,$$

where we substitute $\mathrm{est}^{\gamma}(f) \stackrel{\mathrm{def}}{=} \epsilon$. From Theorem 1 and Lemma 3 of Srebro and Ben-David (2006) we have the following upper bound of the covering number for a family $\mathcal{K} = \{\kappa_1, \ldots, \kappa_p\}$ of kernels bounded by $R^2 \geq \kappa(x, x)$ and any $\alpha < 1$:

$$\mathcal{N}_m(\mathcal{F}_{\mathcal{K}}, \alpha) \leq 2 \left(\frac{4 e m^3 R^2}{\alpha^2 p}\right)^p \left(\frac{16 m R^2}{\alpha^2}\right)^{\frac{64 R^2}{\alpha^2} \log\left(\frac{\alpha e m}{8R}\right)}. \quad (2)$$

Hence using ℓ_1 norm regularization for MKL the algorithm will choose a small number of kernels from a set of p base kernels. Therefore, fix $d < p$. Hence, making use of the fact that we have $\binom{p}{d}$ different ways of choosing a combination of d kernels (counting argument), and making a further application of p (union bound) we get:

$$\binom{p}{d} 2\mathcal{N}_{2m}(\mathcal{F}_{\mathcal{K}}, \gamma/2) \exp\left(-\frac{\epsilon^2 m}{8}\right) = \frac{\delta}{p}.$$

Applying (2) to upper bound the covering number and solving for ϵ yields the result. □

The estimation error of Theorem 1 is of the order:

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{p + R^2/\gamma^2}{m}}\right),$$

where $\tilde{\mathcal{O}}$ hides logarithmic factors.[1] Our bound of Theorem 2 is:

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{\log p + R^2/\gamma^2 + 2d}{m}}\right),$$

with only a logarithmic dependency on p, but an extra additive term of twice the number of kernels chosen d. Typically $d \ll p$.

This bound can be smaller than the bound of Srebro and Ben-David (2006) if there are a large number of kernels in the kernel family (possibly exponentially large) but only a sparse (small) number of kernels chosen in the final combination. A recent algorithm has this property, where a small number of kernels are chosen from an exponentially large set of base kernels (Bach, 2009). The number of kernels used in the experiments of Bach (2009) were of the order of more than $p \geq 10^{30}$, but the algorithm chose a much smaller number of kernels $d \approx 300$ in the final solution. Clearly, in this case we can expect Theorem 2 to give a significantly tighter bound than Theorem 1.

Figure 1 displays bound plots for Theorem 1 and Theorem 2. We plot the estimation error of these bounds (*i.e.*, $\text{est}^{\gamma}(f)$ in both Theorems) as a function of m and p, and choose $d = 300$ for our proposed bound (Theorem 2). The other parameters of the bound were set to $\gamma/R = 0.2$ and $\delta = 0.01$ (using the setup of Cortes et al. (2010)). Our bound is depicted with solid lines and is clearly tighter than Theorem 1 and not effected so much with varying values of p – this is because we only have a logarithmic dependency on p. Hence, all three plots of our bound (for varying values of p) have similar values, with the curves being very close together. Also, when $p = m$ the bound of Theorem 1 becomes uninformative as very early on it starts to increase with an increase in the number of examples m, whereas our bound decreases with the number of training examples m. Hence, our bound is still be informative in this case.

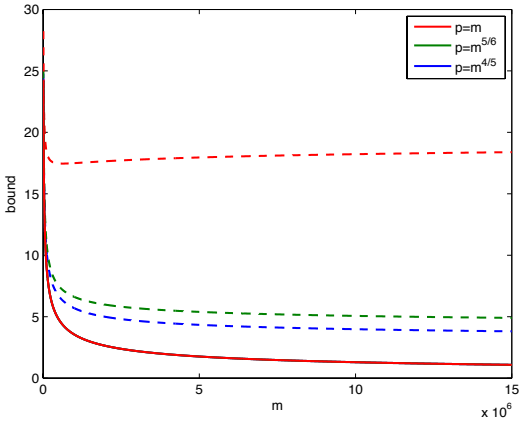

Figure 1: Bound plot comparing the bound of Srebro and Ben-David (2006) and our sample-compressed bound for a normalized margin $\gamma/R = 0.2$ and $\delta = 0.01$. Theorem 1 is given by the dashed lines, and Theorem 2 by the solid lines. Our bound is plotted for $d = 300$ kernels chosen from p base kernels.

5 ADDITIVE RADEMACHER COMPLEXITY BOUND FOR MKL

In this section we derive a novel Rademacher complexity bound for MKL (Ying and Campbell, 2009; Cortes et al., 2010) which does not have a *multiplicative* interaction between the margin complexity term and the dimensionality of the kernel family. In Srebro and Ben-David (2006) they state that it *may* not be possible to have *additive* behavior for Rademacher bounds for MKL. However, we show that it is possible for the case of convex combinations of base kernels by using a result for Rademacher complexities of convex hulls. We begin by stating the following well-known concentration inequality, followed by a definition of Rademacher complexity.

Theorem 3 (McDiarmid, 1989)**.** *Let $X_1, \ldots, X_m$ be independent random variables taking values in a set A, and assume that $f : A^m \mapsto \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_m,\hat{x}_i \in A} |f(x_1,\ldots,x_m) - f(x_1,\ldots,\hat{x}_i,x_{i+1},\ldots,x_m)| \leq c_i, 1 \leq i \leq m.$$

Then for all $\epsilon > 0$

$$\Pr\{f(X_1,\ldots,X_m) - \mathbb{E}f(X_1,\ldots,X_m) \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

Definition 3 (Rademacher complexity)**.** *For a sample $\mathbf{x} = \{x_1,\ldots,x_m\}$ generated by a distribution $\mathcal{D}_{\mathcal{X}}$ on a set $\mathcal{X}$ and a real-valued function class $\mathcal{F}$ with domain $\mathcal{X}$, the* empirical Rademacher complexity *of $\mathcal{F}$ is the random variable*

$$\hat{R}_m(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{2}{m}\sum_{i=1}^m \sigma_i f(x_i) \mid x_1,\ldots,x_m\right].$$

where $\sigma = (\sigma_1,\ldots,\sigma_m)$ are independent uniform $\{\pm 1\}$-valued (Rademacher) random variables. The

[1] The exact form of the bound with all logarithmic factors and constants can be found in $\text{est}^{\gamma}(f)$ of Theorem 1. Similarly for Theorem 2.

(true) Rademacher complexity *is:*

$$R_m(\mathcal{F}) = \mathbb{E}_{\mathbf{x}}\left[\hat{R}_m(\mathcal{F})\right] = \mathbb{E}_{\mathbf{x}\sigma}\left[\sup_{f\in\mathcal{F}}\frac{2}{m}\sum_{i=1}^{m}\sigma_i f(x_i)\right].$$

The standard margin-based Rademacher bound for learning theory is given in the following theorem.

Theorem 4 (Bartlett and Mendelson, 2002)**.** *Fix $\gamma > 0$ and $\delta \in (0,1)$, and let $\mathcal{F}$ be a class of functions mapping from $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$ to $[0,1]$. Let $\mathbf{z} = \{z_i\}_{i=1}^m$ be drawn independently according to a probability distribution $\mathcal{D}$. Then with probability $1-\delta$ over random draws of samples of size m, every $f \in \mathcal{F}$ satisfies*

$$\text{err}(f) \quad \leq \quad \hat{\text{err}}^{\gamma}(f) + \frac{1}{\gamma}\hat{R}_m(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

This bound is quite general and applicable to various learning algorithms if an *empirical Rademacher complexity* $\hat{R}_m(\mathcal{F})$ of the function class $\mathcal{F}$ can be found efficiently. For kernel method algorithms a well-known result uses the trace of the kernel matrix to bound the empirical Rademacher complexity.

Theorem 5 (Bartlett and Mendelson, 2002)**.** *If $\kappa : \mathcal{X}\times\mathcal{X} \mapsto \mathbb{R}$ is a kernel, and $\mathbf{x} = \{x_1,\ldots,x_m\}$ is a sample of points from $\mathcal{X}$, then the empirical Rademacher complexity of the class $\mathcal{F}(B)$ with bounded norm $\|w\|_2 \leq B$ satisfies*

$$\hat{R}_m(\mathcal{F}(B)) \leq \frac{2B}{m}\sqrt{\sum_{i=1}^{m}\kappa(x_i,x_i)}.$$

Furthermore, if $R^2 \geq \kappa(x,x)$ for all $x \in \mathcal{X}$ and κ is a normalized kernel such that $\sum_{i=1}^m \kappa(x_i,x_i) = m$ then we have:

$$\frac{2B}{m}\sqrt{\sum_{i=1}^{m}\kappa(x_i,x_i)} \leq 2B\sqrt{\frac{R^2}{m}}.$$

The problem of learning kernels from a convex combination of base kernels can be viewed as a convex hull:

$$\text{conv}_B(\mathcal{F}) = \left\{\sum a_i f_i : f_i \in \mathcal{F}, a_i \in \mathbb{R} \geq 0, \sum a_i \leq B\right\} \quad (3)$$

We are interested in the empirical Rademacher complexity of a convex hull as given by Equation (3), given in the following result.

Theorem 6 (Ambroladze and Shawe-Taylor, 2004)**.** *The empirical Rademacher complexity of the convex hull $\text{conv}_B(\mathcal{F})$ of function class $\mathcal{F}$ satisfies*

$$\hat{R}_m(\text{conv}_B(\mathcal{F})) \leq B\hat{R}_m(\mathcal{F}).$$

Given all of the results from above, we are now in a position to state the following theorem, which proves an upper bound for the empirical Rademacher complexity of the joint function class $\mathcal{F}_{\mathcal{K}}$.

Theorem 7. *Let $\mathbf{x} = \{x_1,\ldots,x_m\}$ be an m-sample of points from $\mathcal{X}$, then the empirical Rademacher complexity $\hat{R}_m$ of the class $\mathcal{F}_{\mathcal{K}} = \cup_{\kappa_j\in\mathcal{K}}\mathcal{F}_j$, $j \in \{1,\ldots,p\}$, satisfies:*

$$\hat{R}_m(\mathcal{F}_{\mathcal{K}}) \leq \max_{1\leq j\leq p}\hat{R}_m(\mathcal{F}_j) + 8\sqrt{\frac{\ln((p+1)/\delta)}{2m}}.$$

Proof. Let σ^* be a realization of a Rademacher sequence and recall that for a family of kernels $\mathcal{K} = \{\kappa_1,\ldots,\kappa_p\}$ we have the following function class $\mathcal{F}_{\mathcal{K}} = \cup_{\kappa_j\in\mathcal{K}}\mathcal{F}_j$, for all $j \in \{1,\ldots,p\}$. Then with probability at least $1-\delta$ over the generation of this sequence we know that:

$$\begin{aligned}
\hat{R}_m(\mathcal{F}_{\mathcal{K}}) &= \mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}_{\mathcal{K}}}\frac{2}{m}\sum_{i=1}^{m}\sigma_i f(x_i)\right] \\
&\leq \left[\sup_{f\in\mathcal{F}_{\mathcal{K}}}\frac{2}{m}\sum_{i=1}^{m}\sigma_i^* f(x_i)\right] \\
&\qquad +4\sqrt{\frac{\ln((p+1)/\delta)}{2m}} \\
&\leq \max_{1\leq j\leq p}\left[\sup_{f\in\mathcal{F}_j}\frac{2}{m}\sum_{i=1}^{m}\sigma_i^* f(x_i)\right] \\
&\qquad +4\sqrt{\frac{\ln((p+1)/\delta)}{2m}} \\
&\leq \max_{1\leq j\leq p}\mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}_j}\frac{2}{m}\sum_{i=1}^{m}\sigma_i f(x_i)\right] \\
&\qquad +8\sqrt{\frac{\ln((p+1)/\delta)}{2m}} \\
&= \max_{1\leq j\leq p}\hat{R}_m(\mathcal{F}_j) + 8\sqrt{\frac{\ln((p+1)/\delta)}{2m}},
\end{aligned}$$

where the second line follows from an application of Theorem 3, the third line by observing that the supremum of a joint function class (*i.e.*, $\cup\mathcal{F}_j$) will always be upper bounded by the maximum function in one of the function classes, the next line by taking the expectation over σ to get the final line in terms of the empirical Rademacher complexity of a single function class $\mathcal{F}_j$. □

Therefore we have the following generalization error bound for MKL in the case of a convex combination of kernels.

Theorem 8. *Fix $\gamma > 0$ and $\delta \in (0,1)$. Let $\mathcal{K} = \{\kappa_1,\ldots,\kappa_p\}$ be a family of kernels containing p base kernels and let $\mathbf{z} = \{z_i\}_{i=1}^m$ be a randomly generated*

sample from distribution $\mathcal{D}$. Then with probability $1-\delta$ over random draws of samples of size m, every $f \in \mathcal{F}_{\mathcal{K}_{\text{con}}}$ satisfies

$$\begin{aligned}\text{err}(f) \quad \leq \quad & \hat{\text{err}}^{\gamma}(f) + \frac{2}{\gamma m} \max_{1\leq j\leq p} \sqrt{\sum_{i=1}^{m} \kappa_j(x_i, x_i)} \\ & + 11\sqrt{\frac{\ln(p+3)}{2m} + \frac{\ln(1/\delta)}{2m}}\end{aligned}$$

Also, if each kernel κ_j is normalized and bounded by $R^2 \geq \kappa_j(x,x)$ for all $x \in \mathcal{X}$ and $j \in \{1,\ldots,p\}$, we have:

$$\begin{aligned}\text{err}(f) \quad \leq \quad & \hat{\text{err}}^{\gamma}(f) + 2\sqrt{\frac{R^2/\gamma^2}{m}} \\ & + 11\sqrt{\frac{\ln(p+3)}{2m} + \frac{\ln(1/\delta)}{2m}}.\end{aligned}$$

Proof. We view each feature space $\mathcal{F}_j$ as the space for a new kernel. Hence, we have:

$$\begin{aligned}\text{err}(f) \quad \leq \quad & \hat{\text{err}}^{\gamma}(f) + \frac{1}{\gamma}\hat{R}_m(\mathcal{F}_{\mathcal{K}_{\text{con}}}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \\ \leq \quad & \hat{\text{err}}^{\gamma}(f) + \frac{B}{\gamma}\hat{R}_m(\mathcal{F}_{\mathcal{K}}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \\ \leq \quad & \hat{\text{err}}^{\gamma}(f) + \frac{1}{\gamma}\max_{1\leq j\leq p}\hat{R}_m(\mathcal{F}_j) \\ & + 11\sqrt{\frac{\ln((p+3)/\delta)}{2m}} \\ \leq \quad & \hat{\text{err}}^{\gamma}(f) + \frac{2}{\gamma m} \max_{1\leq j\leq p} \sqrt{\sum_{i=1}^{m} \kappa_j(x_i, x_i)} \\ & + 11\sqrt{\frac{\ln((p+3)/\delta)}{2m}} \\ \leq \quad & \hat{\text{err}}^{\gamma}(f) + 2\sqrt{\frac{R^2/\gamma^2}{m}} \\ & + 11\sqrt{\frac{\ln(p+3)}{2m} + \frac{\ln(1/\delta)}{2m}}.\end{aligned}$$

Where the first line is given by Theorem 4, the second line comes from applying Theorem 6, the third by applying Theorem 7 and considering that the MKL formulation is a convex hull (see Equation (3)) such that $B \leq 1$. The fourth line is obtained by applying the first inequality in Theorem 5. The final line is obtained by applying the second inequality in Theorem 5 for the case when each kernel κ_j is normalized and bounded by R^2. □

Recently, Cortes et al. (2010) have proposed the following Rademacher bound for MKL:

Theorem 9 (Cortes et al., 2010)**.** *Fix $\gamma > 0$ and $\delta \in (0,1)$. Then, for any $p > 0$ base kernels and $R^2 > \kappa_j(x,x)$ for all $x \in \mathcal{X}$ and all $j \in \{1,\ldots,p\}$, for any $f \in \mathcal{F}_{\mathcal{K}_{\text{con}}}$ with probability $1-\delta$ we have:*

$$\text{err}(f) \quad \leq \quad \hat{\text{err}}^{\gamma}(f) + 2\sqrt{\frac{\eta e\lceil \log p\rceil R^2/\gamma^2}{m}} + 3\sqrt{\frac{\ln(2/\delta)}{2m}},$$

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

The estimation error of Theorem 9 (Cortes et al., 2010) is (ignoring constants):

$$\mathcal{O}\left(\sqrt{\frac{(\log p)R^2/\gamma^2}{m}}\right),$$

which contains a multiplicative dependence between $\log p$ and γ. Our bound is:

$$\mathcal{O}\left(\sqrt{\frac{\ln p + R^2/\gamma^2}{m}}\right).$$

This is *additive* in $\ln p$ and the margin (complexity) term γ. Therefore, we have an additive expression similar to Srebro and Ben-David (2006) and a logarithmic dependence for the kernel complexity term, similar to Cortes et al. (2010).

Figure 2 shows bound plots of the estimation error (*i.e.*, est^{γ}) for our bound of Theorem 8 and the bound of Theorem 9. Figure 2 shows the plots using the same setup as that given in Cortes et al. (2010), namely a normalized margin $\gamma/R = 0.2$ and $\delta = 0.01$. The bounds are plotted as a function of m and different values of p. It is clear from the plots on the right that our bound of Theorem 8 is tighter than that of Theorem 9. Furthermore, we carried out some experiments using standard SVM software and benchmark datasets to gauge the size of the margins typically encountered after training. The normalized margins we found were closer to 0.02 than 0.2, so Figure 3 shows the same bound plots but for normalized margin $\gamma/R = 0.02$. It is clear that our bound is still tighter than the bound of Cortes et al. (2010). Furthermore, we can see from both plots that the change in the number of base kernels does not alter the value of our bounds as much as the change encountered for Theorem 9. This is because we only have an additive dependency between $\ln p$ and the margin γ.

6 CONCLUSION

We proposed two novel bounds for MKL. The first applies a simple counting analysis plus union bound over a previously published MKL bound. The idea is to count the number of ways of choosing the kernels included in the final combination and to compute

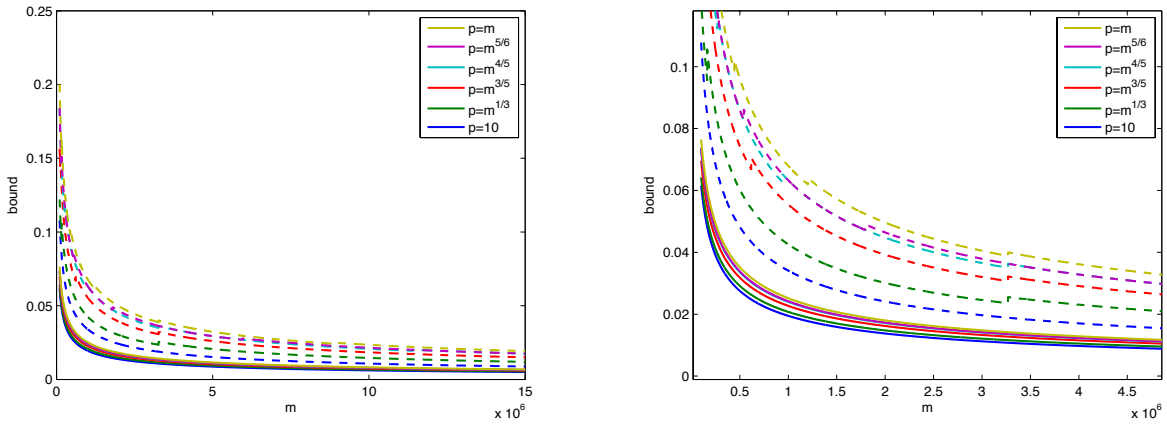

Figure 2: Bound plot comparing the bound of Cortes et al. (2010) and our bound for a normalized margin $\gamma/R = 0.2$ and $\delta = 0.01$. The bound of Theorem 9 Cortes et al. (2010) is given by the dashed lines, and our bound of Theorem 8 by solid lines. The plot on the right is a zoomed in version of the plot on the left.

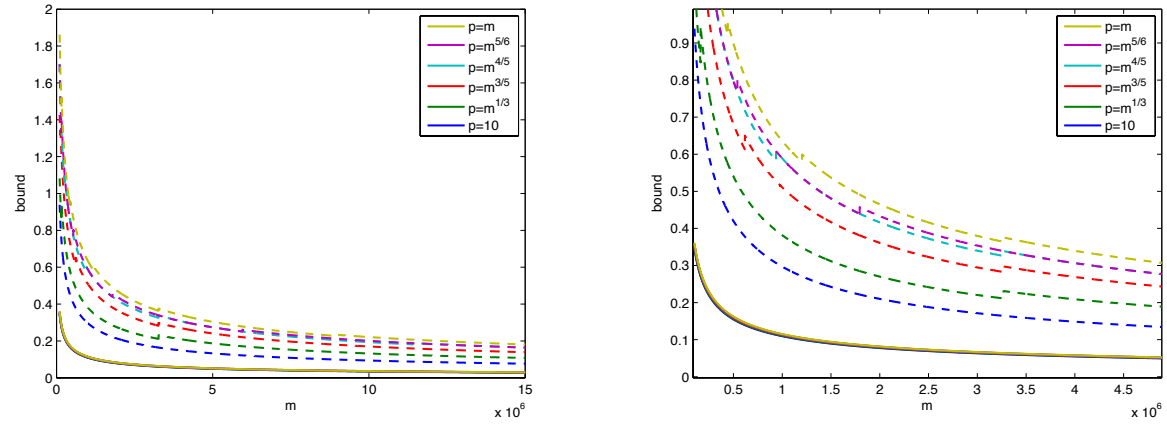

Figure 3: Bound plot comparing the bound of Cortes et al. (2010) and our bound for a normalized margin $\gamma/R = 0.02$ and $\delta = 0.01$. The bound of Theorem 9 is given by the dashed lines, and our bound of Theorem 8 by solid lines. The plot on the right is a zoomed in version of the plot on the left.

a union bound over all of these choices. Hence, we can apply this technique over the bound of Srebro and Ben-David (2006) for the choice of kernels. We show empirically that this bound is tighter than the Srebro and Ben-David (2006) result when a small number of base kernels is chosen from a large kernel family; algorithms such as that proposed by Bach (2009) have this desirable property. Hence, our bound corroborates the impressive empirical results obtained by Bach (2009), and suggests that we will have good generalization whenever a sparse set of kernels can be found from a very large family of kernels.

The second bound is, to our knowledge, the first MKL bound using Rademacher complexity which is additive in the kernel complexity and margin term. It uses Rademacher theory results from the boosting literature, and is tighter than all previously published MKL bounds for learning a convex combination of kernels, including the recent bound of Cortes et al. (2010) – which is also a Rademacher complexity bound but with a multiplicative interaction between the kernel complexity and margin terms.

The Rademacher bound motivates an LPBoost (Demiriz et al., 2002) framework for multiple kernel learning. We may view the set of norm bounded linear functions in each kernel's feature space as weak learners and hence apply the LPBoost algorithm in this case. This would correspond to a 1-norm regularization of the choice of kernels and result in a different algorithm for SimpleMKL (Rakotomamonjy et al., 2008). The

details of this algorithm will be presented in a longer format of the current paper.

Finally it should be noted that although the bounds of this paper and the bound of Cortes et al. (2010) are tighter than the Srebro and Ben-David (2006) bound, they are in fact less general. The bound of Srebro and Ben-David (2006) can be applied in the case of a non-linear combination of kernels, or a parameterized class of kernels. However, the bounds we presented (and also Cortes et al. (2010)) are for the regime of MKL that is considered the most popular in the literature – where we have a finite family of kernels to choose a linear or convex combination from. Therefore, it would be an interesting research direction to apply the bounds of this paper to a more general class of kernel families.

Acknowledgements

We thank Xingxin Xu for pointing out typos in Theorem 7 . The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under *grant agreement* no. 216529, Personal Information Navigator Adapting Through Viewing, PinView, and in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

References

M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821 – 837, 1964.

A. Ambroladze and J. Shawe-Taylor. Complexity of pattern classes and lipschitz property. In *Algorithmic Learning Theory*, volume 3244 of *Lecture Notes in Computer Science*, pages 181–193. Springer, 2004.

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

A. Argyriou, C. A. Micchelli, and M. Pontil. Learning convex combinations of continuously parameterized basic kernels. In *Computational Learning Theory*, volume 3559 of *Lecture Notes in Computer Science*, pages 338–352. Springer, 2005.

F. R. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 105–112. 2009.

F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6, New York, NY, USA, 2004. ACM.

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

C. Cortes, M. Mohri, and A. Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the twenty-seventh international conference on Machine learning*, New York, NY, USA, 2010. ACM.

A. Demiriz, K. P. Bennett, and J. Shawe-Taylor. Linear programming boosting via column generation. *Machine Learning*, 46(13):225–254, 2002.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1): 1–50, 2002.

G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

C. McDiarmid. On the method of bounded differences. In . L. M. S. L. N. Series, editor, *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge, 1989.

C. A. Micchelli, M. Pontil, Q. Wu, and D.-X. Zhou. Error bounds for learning the kernel. Technical report, University College London, UK, 2005.

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

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, U.K., 2004.

N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *Computational Learning Theory*, volume 4005 of *Lecture Notes in Computer Science*, pages 169–183. Springer, 2006.

Y. Ying and C. Campbell. Generalization Bounds for Learning the Kernel. In *Proceedings of the 22nd Annual Conference on Computational Learning Theory*. Springer, Berlin, 2009.

Y. Ying and D.-X. Zhou. Learnability of gaussians with flexible variances. *Journal of Machne Learning Research*, 8:249–276, May 2007. ISSN 1532-4435.