A Note on Improved Loss Bounds for Multiple Kernel Learning

Zakria Hussain & John Shawe-Taylor
Department of Computer Science
University College London
London, WC1E 6BT, UK
e-mail: {z.hussain,jst}@cs.ucl.ac.uk

Mario Marchand
Computer Science and Software Engineering
Laval University, Québec (QC), Canada
e-mail: mario.marchand@ift.ulaval.ca

May 13, 2014

Abstract

In this paper, we correct an upper bound, presented in [4], on the generalisation error of classifiers learned through multiple kernel learning. The bound in [4] uses Rademacher complexity and has an *additive* dependence on the logarithm of the number of kernels and the margin achieved by the classifier. However, there are some errors in parts of the proof which are corrected in this paper. Unfortunately, the final result turns out to be a risk bound which has a *multiplicative* dependence on the logarithm of the number of kernels and the margin achieved by the classifier.

1 Introduction

We refer to [4] for the motivation and definitions of multiple kernel learning. It presents a number of results, including a new Rademacher complexity bound on the generalisation error of classifiers learned from a multiple kernel class with a logarithmic dependence on the number of kernels used and with that logarithm entering additively into the bound—that is, independently

1

of the complexity of the individual kernels or the margin of the classifier on the training set.

In this paper, we follow the approach presented in [4] but correct some of the errors that are present. Unfortunately, the Rademacher complexity risk bound turns out to exhibit a *multiplicative* dependence on the logarithm of the number of kernels and the margin achieved by the classifier.

2   Detailed proof

2.1   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 ([1]). *A kernel is a function κ that for all $x, x' \in \mathcal{X}$ satisfies*

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

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 [7, 8] make use of the $m \times m$ kernel matrix $K = [\kappa(x_i, x_{i'})]_{i,i'=1}^m$ 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:

$$f(x) = \sum_{i=1}^m \alpha_i y_i \kappa(x_i, x) = \langle w, \phi(x) \rangle,$$

where $\alpha = (\alpha_1, \ldots, \alpha_m)$ is the dual weight vector and the corresponding norm of the weight vector is

$$\|w\|^2 = \sum_{i,j=1}^m \alpha_i y_i \alpha_j y_j \kappa(x_i, x_j).$$

Given a kernel κ, we will use $\phi_\kappa(\cdot)$ to denote a feature space mapping satisfying

$$\kappa(x, x') = \langle \phi_\kappa(x), \phi_\kappa(x') \rangle.$$

Hence, learning with a kernel κ can be described as finding a function from the class of functions [9]

$$\mathcal{F}_\kappa = \{x \mapsto \langle w, \phi_\kappa(x) \rangle \mid \|w\|_2 \le 1, \}$$

2

minimising the empirical average of the hinge loss

$$h^\gamma(yf(x)) = \max\left(1 - \frac{yf(x)}{\gamma}, 0\right).$$

where we call $\gamma \in [0, 1]$ the *margin*. For multiple kernel learning we consider a family of kernels $\mathcal{K}$ and the corresponding function class

$$\mathcal{F}_{\mathcal{K}} = \{x \mapsto \langle w, \phi_\kappa(x)\rangle \mid \|w\|_2 \leq 1, \text{ for some } \kappa \in \mathcal{K}\} .$$

For a distribution $\mathcal{D}$, we use the notation $\mathbb{E}_{\mathcal{D}}[f(x)]$ to denote the expected value of $f(x)$ when $x \sim \mathcal{D}$. Given a training set $\mathbf{x}$ we denote $\hat{\mathbb{E}}[f]$ to denote its empirical average over the sample $\mathbf{x}$.

For the generalisation 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:

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

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

$$\text{e}\hat{\text{r}}\text{r}^\gamma(f) = \frac{1}{m}\sum_{i=1}^m \mathbb{I}(y_i f(x_i) < \gamma) = \hat{\mathbb{E}}[\mathbb{I}(y_i f(x_i) < \gamma)],$$

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

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

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

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

We further consider the clipped hinge function:

$$\mathcal{A}^\gamma(s) = \begin{cases} 0; & \text{if } s \geq \gamma \\ 1 - s/\gamma; & \text{if } 0 \leq s \leq \gamma; \\ 1; & \text{otherwise,} \end{cases}$$

and its empirical estimation $\hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))]$. Note that $\text{err}(f) \leq \mathbb{E}_{\mathcal{D}}[\mathcal{A}^\gamma(yf(x))]$, $\hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] \leq \text{e}\hat{\text{r}}\text{r}^\gamma(f)$ and $\hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] \leq \hat{\mathbb{E}}[h^\gamma(yf(x))]$.

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

$$\mathcal{K}_{\mathrm{con}}(\kappa_1, \ldots, \kappa_p) = \left\{ \kappa^\lambda = \sum_{j=1}^p \lambda_j \kappa_j \mid \lambda_j \geq 0, \sum_{j=1}^p \lambda_j = 1 \right\}.$$

Note, p is the complexity of the kernel family (*i.e.*, cardinality of the set of base kernels).

2.2 Rademacher complexity bound for MKL

In this section we correct the MKL risk bound of [4]. We begin by the following definition of Rademacher complexity.

Definition 2 (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 (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 Rademacher bound for function classes is given in the following theorem.

Theorem 1 ([3]). *Fix $\delta \in (0,1)$, 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*

$$\mathbb{E}_\mathcal{D}(f) \quad \leq \quad \hat{\mathbb{E}}(f) + \hat{R}_m(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

4

We have attributed this bound to [3], though, strictly speaking, they used the slightly weaker version of Rademacher complexity including an absolute value of the sum. This version is obtained by a slight tightening of the argument. This bound is quite general and applicable to various learning algorithms if a tight upper bound of *empirical Rademacher complexity* $\hat{R}_m(\mathcal{F})$ of the function class $\mathcal{F}$ can be found. For kernel methods, a well-known result uses the trace of the kernel matrix to bound the empirical Rademacher complexity.

Theorem 2 ([3]). *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 $\mathcal{F}_\kappa$ satisfies*

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

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

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

The problem of learning kernels from a convex combination of base kernels is related to using the convex hull of a set of functions. Consider

$$\mathrm{con}(\mathcal{F}) = \left\{ \sum_j a_j f_j \mid f_j \in \mathcal{F}, a_j \geq 0, \sum_j a_j \leq 1 \right\}. \tag{2}$$

Since adding kernels corresponds to concatenating feature spaces, it is clear that (here w_j is the restriction of w to the feature space defined by the mapping $\phi_{\kappa_j}(\cdot)$ corresponding to kernel κ_j)

$$
\begin{aligned}
\mathcal{F}_{\mathcal{K}_{\mathrm{con}}(\kappa_1,\ldots,\kappa_p)} &= \left\{ x \mapsto \langle w, \phi_\kappa(x) \rangle \mid \|w\|_2 \leq 1, \kappa = \sum_{j=1}^{p} \lambda_j \kappa_j, \sum_{j=1}^{p} \lambda_j = 1 \right\} \\
&= \left\{ x \mapsto \sum_{j=1}^{p} \sqrt{\lambda_j} \|w_j\| \left\langle \frac{w_j}{\|w_j\|}, \phi_{\kappa_j}(x) \right\rangle \mid \|w\|_2 \leq 1, \sum_{j=1}^{p} \lambda_j = 1 \right\} \\
&= \mathrm{con}\left(\bigcup_{j=1}^{p} \mathcal{F}_{\kappa_j} \right), \tag{3}
\end{aligned}
$$

since, by the Cauchy Schwartz inequality, we have

$$\sum_{j=1}^{p} \sqrt{\lambda_j} \|w_j\| \leq \sqrt{\sum_{j=1}^{p} \lambda_j} \sqrt{\sum_{j=1}^{p} \|w_j\|^2} \leq 1.$$

Hence, we are interested in the empirical Rademacher complexity of a convex hull as given by Equation (2), which is well known to satisfy

$$\hat{R}_m(\text{con}(\mathcal{F})) = \hat{R}_m(\mathcal{F}). \tag{4}$$

Furthermore, following [5] and [2], we have the following result.

Theorem 3 ([5])**.** *The empirical Rademacher complexity of the function class $\mathcal{L}(\mathcal{F})$ where $\mathcal{L}(\cdot)$ is Lipschitz function with Lipschitz constant L is bounded by*

$$\hat{R}_m(\mathcal{L}(\mathcal{F})) \leq L\hat{R}_m(\mathcal{F}).$$

Given all these results, we are now in a position to state the following theorem, which proves a high probability upper bound for the empirical Rademacher complexity of a union of function classes $\bigcup_{j=1}^{p} \mathcal{F}_j = \mathcal{F}$.

Theorem 4. *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 $\mathcal{F} = \cup_{j=1}^{p} \mathcal{F}_j$, where the range of all the functions in $\mathcal{F}$ is $[0, 1]$, satisfies:*

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

Proof. Since $\mathcal{F}$ is the union of p function classes, we have

$$\hat{R}_m(\mathcal{F}) = \mathbb{E}_\sigma \max_{1 \leq j \leq p} \sup_{f \in \mathcal{F}_j} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i).$$

From Jensen's inequality, we have, for any $\lambda \geq 0$, that

$$\exp\left(\lambda \hat{R}_m(\mathcal{F})\right) \leq \mathbb{E}_\sigma \exp\left(\lambda \left[\max_{1 \leq j \leq p} \sup_{f \in \mathcal{F}_j} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i)\right]\right)$$

$$= \mathbb{E}_\sigma \max_{1 \leq j \leq p} \exp\left(\lambda \left[\sup_{f \in \mathcal{F}_j} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i)\right]\right)$$

$$\leq \sum_{j=1}^{p} \mathbb{E}_\sigma \exp\left(\lambda \left[\sup_{f \in \mathcal{F}_j} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i)\right]\right). \tag{5}$$

6

Now, for any fixed function class $\mathcal{F}_j$ and any fixed training sample, let

$$\xi(\sigma_1, \ldots, \sigma_m) \stackrel{\text{def}}{=} \sup_{f \in \mathcal{F}_j} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \,.$$

A basic result of McDiarmid [6] states that for any $\lambda \geq 0$, we have

$$\mathbb{E}e^{\lambda\xi} \leq e^{\frac{\lambda^2}{8} \sum_{i=1}^{m} c_i^2} \cdot e^{\lambda\mathbb{E}\xi} \,,$$

where, for all i, we have

$$\sup_{\sigma_1, \ldots, \sigma_m, \hat{\sigma}_i} |\xi(\sigma_1, \ldots, \sigma_m) - \xi(\sigma_1, \ldots, \sigma_{i-1}, \hat{\sigma}_i, \sigma_{i+1}, \ldots, \sigma_m)| \leq c_i \,.$$

In our case, we have that $c_i \leq 4/m \; \forall i \in \{1, \ldots, m\}$. Hence, from Equation (5), we have

$$\exp\left(\lambda \hat{R}_m(\mathcal{F})\right) \leq e^{2\lambda^2/m} \sum_{j=1}^{p} e^{\lambda \hat{R}_m(\mathcal{F}_j)} \,.$$

By taking the logarithm on both sides of this equation, we obtain

$$\begin{aligned}
\lambda \hat{R}_m(\mathcal{F}) &\leq \frac{2\lambda^2}{m} + \ln\left[\sum_{j=1}^{p} e^{\lambda \hat{R}_m(\mathcal{F}_j)}\right] \\
&\leq \frac{2\lambda^2}{m} + \ln\left[p \cdot \max_{1 \leq j \leq p} e^{\lambda \hat{R}_m(\mathcal{F}_j)}\right] \\
&\leq \frac{2\lambda^2}{m} + \ln(p) + \max_{1 \leq j \leq p} \lambda \hat{R}_m(\mathcal{F}_j) \,.
\end{aligned}$$

Hence, we have

$$\hat{R}_m(\mathcal{F}) \leq \frac{2\lambda}{m} + \frac{1}{\lambda} \ln(p) + \max_{1 \leq j \leq p} \hat{R}_m(\mathcal{F}_j) \,.$$

The theorem then follows from this equation by choosing

$$\lambda = \sqrt{\frac{m}{2} \ln p} \,.$$

$\square$

Recall the function $\mathcal{A}^\gamma(\cdot)$ and the properties $\text{err}(f) \leq \mathbb{E}_\mathcal{D}[\mathcal{A}^\gamma(yf(x))]$ and $\mathbb{E}[\mathcal{A}^\gamma(yf(x))] \leq \text{err}^\gamma(f)$. Therefore we have the following generalization error bound for MKL in the case of a convex combination of kernels.

Theorem 5. *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 the random draws of samples of size m, every $f \in \mathcal{F}_{\mathcal{K}_{\mathrm{con}}}$ satisfies*

$$
\mathrm{err}(f) \;\leq\; \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \frac{1}{\gamma} \left[\max_{1 \leq j \leq p} \frac{2}{m} \sqrt{\sum_{i=1}^m \kappa_j(x_i, x_i)} + \sqrt{\frac{8 \ln p}{m}} \right]
$$
$$
+ 3\sqrt{\frac{\ln(2/\delta)}{2m}} .
$$

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

$$
\mathrm{err}(f) \;\leq\; \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \frac{1}{\gamma} \left[\frac{2R}{\sqrt{m}} + \sqrt{\frac{8 \ln p}{m}} \right] + 3\sqrt{\frac{\ln(2/\delta)}{2m}} .
$$

Proof. Each kernel κ_j defines the class $\mathcal{F}_j = \{x \mapsto \langle w, \phi_{\kappa_j}(x) \rangle : \|w\| \leq 1\}$. Hence, applying Theorem 1 to the class $\mathcal{A}^\gamma(\mathcal{F}_{\mathcal{K}}) = \{\mathcal{A}^\gamma \circ f : f \in \mathcal{F}_{\mathcal{K}}\}$, we have

$$
\begin{aligned}
\mathrm{err}(f) \;\leq\;& \mathbb{E}_{\mathcal{D}}[\mathcal{A}^\gamma(yf(x))] \\[4pt]
\leq\;& \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \hat{R}_m(\mathcal{A}^\gamma(\mathcal{F}_{\mathcal{K}_{\mathrm{con}}})) + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \\[4pt]
\leq\;& \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \frac{1}{\gamma} \hat{R}_m(\mathcal{F}_{\mathcal{K}_{\mathrm{con}}}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \\[4pt]
=\;& \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \frac{1}{\gamma} \hat{R}_m \left(\bigcup_{j=1}^p \mathcal{F}_{\kappa_j} \right) + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \\[4pt]
\leq\;& \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \frac{1}{\gamma} \left[\max_{1 \leq j \leq p} \hat{R}_m(\mathcal{F}_j) + \sqrt{\frac{8 \ln p}{m}} \right] + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \\[4pt]
\leq\;& \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \frac{1}{\gamma} \left[\max_{1 \leq j \leq p} \frac{2}{m} \sqrt{\sum_{i=1}^m \kappa_j(x_i, x_i)} + \sqrt{\frac{8 \ln p}{m}} \right] + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \\[4pt]
\leq\;& \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \frac{1}{\gamma} \left[\max_{1 \leq j \leq p} \frac{2}{m} \sqrt{mR^2} + \sqrt{\frac{8 \ln p}{m}} \right] + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \\[4pt]
=\;& \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \frac{1}{\gamma} \left[\frac{2R}{\sqrt{m}} + \sqrt{\frac{8 \ln p}{m}} \right] + 3\sqrt{\frac{\ln(2/\delta)}{2m}} ,
\end{aligned}
$$

where the third line comes from applying Theorem 3 with with Lipschitz constant $L = 1/\gamma$. The forth line comes by applying Equation (4). The fifth line comes by applying Theorem 4. The 6th line follows from Theorem 2. Finally, the 7th line follows from the hypothesis that $\kappa_j(x,x) \leq R^2 \; \forall x$. $\square$

3 Discussion

Using the notation from above, the un-normalized version of the bound of Theorem 8 of [4] is

$$\text{err}(f) \quad \leq \quad \hat{\text{err}}^\gamma(f) + \frac{2}{\gamma m} \max_{1 \leq j \leq p} \sqrt{\sum_{i=1}^{m} \kappa_j(x_i, x_i)} + 5\sqrt{\frac{\ln((p+3)/\delta)}{2m}}.$$

Comparing this to Theorem 5 (the corrected version), we can see that the major difference is the fact that $1/\gamma$ is multiplying $\ln p$ in Theorem 5, while it is not in the Theorem 8 of [4]. However, the latter was obtained by *incorrectly* assuming that $\hat{R}_m(\mathcal{A}^\gamma(\text{con}(\mathcal{F})))$ is upper bounded by $\hat{R}_m(\text{con}(\mathcal{A}^\gamma(\mathcal{F})))$.

While Theorem 2 of [4] shows an additive dependence on the logarithm of the number of kernels it has an additional term that includes the number of kernels d involved in the final solution and this number is also multiplied by the logarithm of the number of kernels. However, these quantities are separate from the main margin complexity term. A similar result could be obtained for the Rademacher bound given here resulting in a partial independence between the complexity and number of kernel terms, but with the final number of active kernels entering as an additional complexity term.

Acknowldgements

References

[1] 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.

[2] 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.

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

[4] Z. Hussain and J. Shawe-Taylor. Improved loss bounds for multiple kernel learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[5] S. Kakade and A. Tewari. Rademacher composition. *CMSC 35900 Learning Theory*, pages 1–22, 2008.

[6] 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.

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

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

[9] 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.