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

July 1, 2011

Abstract

The paper [5] presented a bound on the generalisation error of classifiers learned through multiple kernel learning. The bound has (an improved) *additive* dependence on the number of kernels (with the same logarithmic dependence on this number). However, parts of the proof were incorrectly presented in that paper. This note remedies this weakness by restating the problem and giving a detailed proof of the Rademacher complexity bound from [5].

1 Introduction

We refer to [5] for the motivation and definitions of multiple kernel learning. The paper [5] presented a number of results including a new 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 of the complexity of the individual kernels or the margin of the classifier on the training set.

An anonymous referee made us aware of a weakness in the presentation of this proof that was subsequently highlighted in a recent tutorial [4]. In this short note we give a detailed proof of the bound presented in [5] albeit with one constant weakened from 5 to 11.

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]:

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

minimising the empirical average of the hinge loss

$$h^\gamma(yf(x)) = \max(\gamma - yf(x), 0),$$

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

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

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:

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

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

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

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

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

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 $\mathrm{est}^{\gamma}(f)$:

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

We further consider the 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 $\mathrm{err}(f) \leq \mathbb{E}_{\mathcal{D}}[\mathcal{A}^{\gamma}(yf(x))]$, $\hat{\mathbb{E}}[\mathcal{A}^{\gamma}(yf(x))] \leq \mathrm{err}^{\gamma}(f)$ 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 Additive Rademacher complexity bound for MKL

In this section we derive our additive Rademacher bound from [5], using considerably more detail. We begin by stating the following well-known concentration inequality, followed by a definition of Rademacher complexity.

Theorem 1 ([6]). *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 $\epsilon > 0$

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

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 $\{\pm1\}$-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 margin-based Rademacher bound for learning theory is given in the following theorem.

Theorem 2 ([3]). *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*

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

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 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 3 ([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)

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

since by Cauchy Schwartz

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

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

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

Furthermore, we have the following result.

Theorem 4 ([2])**.** *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 of the results from above, 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 joint function class $\mathcal{F} = \bigcup_{j=1}^{p} \mathcal{F}_j$.

Theorem 5. *Let $\mathbf{x} = \{x_1, \ldots, x_m\}$ be an m-sample of points from $\mathcal{X}$, then with probability at least $1 - \delta$ 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:*

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

Proof. From McDiarmid's inequality we have the following probabilistic upper bound (for some function $g(\mathbf{x}) = g(x_1, \ldots, x_m)$ satisfying $\sup_{\mathbf{x}, \hat{x}_i} |g(\mathbf{x}) - g(x_1, \ldots, \hat{x}_i, \ldots, x_m)| \leq c$):

$$\Pr\left\{g(X_1, \ldots, X_m) - \mathbb{E}[g(X_1, \ldots, X_m)] > \epsilon\right\} < \exp\left(-\frac{2\epsilon^2}{mc^2}\right) =: \hat{\delta}, \tag{4}$$

and conversely:

$$\Pr\left\{g(X_1, \ldots, X_m) - \mathbb{E}[g(X_1, \ldots, X_m)] \leq \epsilon\right\} \geq 1 - \hat{\delta}. \tag{5}$$

Let us define

$$g_j(\sigma) = +\sup_{f \in \mathcal{F}_j} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i),$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)$, and

$$g_0(\sigma) = -\sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i).$$

Note that for all of these functions $c = 4/m$ so that

$$\hat{\delta} = \exp\left(-\frac{m\epsilon^2}{8}\right)$$

so that solving for ϵ gives

$$\epsilon = \sqrt{\frac{8\ln(1/\hat{\delta})}{m}}.$$

We would like to upper bound the following empirical Rademacher complexity,

$$\hat{R}_m(\mathcal{F}) \quad = \quad \mathbb{E}_\sigma\left[\sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \middle| \, x_1, \ldots, x_m\right]. \tag{6}$$

We will ignore the conditioning from now on. From Equation (5) with probability $1 - \hat{\delta}$ we can upper bound the expectation $\mathbb{E}_\sigma[g_0(\sigma)]$ by:

$$\begin{aligned} -\mathbb{E}_\sigma[g_0(\sigma)] \quad &\leq \quad -g_0(\sigma^*) + \epsilon, \\ \implies \mathbb{E}_\sigma\left[\sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i)\right] \quad &\leq \quad \left[\sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i^* f(x_i)\right] + \epsilon, \end{aligned}$$

where $\sigma^* = (\sigma_1^*, \ldots, \sigma_m^*)$ is a realisation of a Rademacher sequence. This 'trick' allows us to remove the expectation. We know that the supremum of a joint function class can be upper bounded by the max over the supremum of each of the function classes. Hence, with probability at least $1 - \hat{\delta}$ this gives us:

$$\begin{aligned} \mathbb{E}_\sigma\left[\sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i)\right] \quad &\leq \quad \max_{1 \leq j \leq p} \left[\sup_{f \in \mathcal{F}_j} \frac{2}{m} \sum_{i=1}^{m} \sigma_i^* f(x_i)\right] + \epsilon, \qquad (7) \\ &= \quad \max_{1 \leq j \leq p} g_j(\sigma^*) + \epsilon. \end{aligned}$$

Next we can make another application of Equation (5) to have a bound in terms of g_j. Using the union bound we have with probability $1 - p\hat{\delta}$ for

$1 \leq j \leq p$:

$$\begin{aligned} g_j(\sigma^*) &\leq \mathbb{E}_\sigma[g_j(\sigma)] + \epsilon, \\ \Longrightarrow \left[\sup_{f\in\mathcal{F}_j} \frac{2}{m}\sum_{i=1}^m \sigma_i^* f(x_i)\right] &\leq \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_j} \frac{2}{m}\sum_{i=1}^m \sigma_i f(x_i)\right] + \epsilon. \qquad (8) \end{aligned}$$

Therefore substituting Equation (8) into Equation (7) gives us with probability $1-(p+1)\hat{\delta}$ an upper bound on the empirical Rademacher complexity of a joint function class:

$$\begin{aligned} \hat{R}_m(\mathcal{F}) &\leq \max_{1\leq j\leq p} \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_j} \frac{2}{m}\sum_{i=1}^m \sigma_i f(x_i)\right] + 2\epsilon. \\ &= \max_{1\leq j\leq p} \hat{R}_m(\mathcal{F}_j) + 2\epsilon. \qquad (9) \end{aligned}$$

Finally substituting $\hat{\delta} = \delta/(p+1)$ gives

$$\epsilon = 4\sqrt{\frac{\log((p+1)/\delta)}{2m}}.$$

Hence, with probability $1-\delta$ we have the final result:

$$\hat{R}_m(\mathcal{F}) \leq \max_{1\leq j\leq p} \hat{R}_m(\mathcal{F}_j) + 8\sqrt{\frac{\log((p+1)/\delta)}{2m}}. \qquad (10)$$

□

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 generalisation error bound for MKL in the case of a convex combination of kernels.

Theorem 6. *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*

$$\text{err}(f) \leq \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \frac{2}{\gamma m}\max_{1\leq j\leq p}\sqrt{\sum_{i=1}^m \kappa_j(x_i,x_i)} + 11\sqrt{\frac{\ln((p+3)/\delta)}{2m}}$$

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:

$$\text{err}(f) \leq \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + 2\sqrt{\frac{R^2/\gamma^2}{m}} + 11\sqrt{\frac{\ln((p+3)/\delta)}{2m}}.$$

Proof. Each kernel κ_j defines a function class

$$\mathcal{F}_j = \{x \mapsto \langle w, \phi_{\kappa_j}(x)\rangle : \|w\| \leq 1\}.$$

Hence, applying Theorem 2 to the class of functions

$$\mathcal{A}^\gamma(\mathcal{F}_\mathcal{K}) = \{\mathcal{A}^\gamma \circ f : f \in \mathcal{F}_\mathcal{K}\},$$

we have:

$$\begin{aligned}
\mathrm{err}(f) &\leq \mathbb{E}_\mathcal{D}[\mathcal{A}^\gamma(yf(x))] \\
&\leq \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \hat{R}_m(\mathcal{A}^\gamma(\mathcal{F}_{\mathcal{K}_{\mathrm{con}}})) + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \\
&\leq \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \max_{1\leq j\leq p} \hat{R}_m(\mathcal{A}^\gamma(\mathcal{F}_j)) + 11\sqrt{\frac{\ln((p+3)/\delta)}{2m}} \\
&\leq \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + \frac{2}{\gamma m}\max_{1\leq j\leq p}\sqrt{\sum_{i=1}^m \kappa_j(x_i,x_i)} + 11\sqrt{\frac{\ln((p+3)/\delta)}{2m}} \\
&\leq \hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))] + 2\sqrt{\frac{R^2/\gamma^2}{m}} + 11\sqrt{\frac{\ln((p+3)/\delta)}{2m}},
\end{aligned}$$

where the second line is given by Theorem 2 with $2\delta/(p+3)$ in place of δ, the third line comes from applying Theorem 5 with $(p+1)\delta/(p+3)$ in place of δ, while the fourth uses a combination of Theorem 4 (note that $\mathcal{A}^\gamma(\cdot)$ is Lipschitz with Lipschitz constant $L = 1/\gamma$) and the first inequality in Theorem 3. The final line is obtained by applying the second inequality in Theorem 3 for the case when each kernel κ_j is bounded by R^2. □

3 Discussion

Theorem 8 presented in our AISTATS paper [5] is (using all the notation from above and only presenting the unnormalised version):

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

Comparing this to Theorem 6, we can see the only differences are:

- A constant of 11 instead of 5.
- Using the tighter empirical loss $\hat{\mathbb{E}}[\mathcal{A}^\gamma(yf(x))]$ as opposed to $\hat{\mathrm{err}}^\gamma(f)$.

It is important to note that the bound is additive in the sense that it combines the logarithm of the number of kernels and the margin additively. Furthermore for this scenario it is tighter than any previously published bounds in the ℓ_1 norm regularised MKL literature.

Acknowledgements

We thank an anonymous reviewer from the Journal of Machine Learning Research who pointed out the flaw in the proof of Theorem 6. We would also like to thank Xinxing Xu for pointing out typos in Theorem 5 which helped correct the constant.

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] C. Cortes, M. Mohri, and A. Rostamizadeh. ICML 2011 tutorial - learning kernels, June 2011.

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

[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.