Generalization Guarantees for a Binary Classification Framework for Two-Stage Multiple Kernel Learning

Purushottam Kar
Department of Computer Science and Engineering
IIT Kanpur
purushot@cse.iitk.ac.in

February 2, 2013

Abstract

We present generalization bounds for the TS-MKL framework for two stage multiple kernel learning. We also present bounds for sparse kernel learning formulations within the TS-MKL framework.

1 Introduction

Recently Kumar *et al* [6] proposed a framework for two-stage multiple kernel learning that combines the idea of target kernel alignment and the notion of a *good* kernel proposed in [1] to learn a good Mercer kernel. More specifically, given a finite set of base kernels $K_1, \ldots, K_p$ over some common domain $\mathcal{X}$, we wish to find some combination of these base kernels that is well suited to the learning task at hand. The paper considers learning a positive linear combination of the kernels $K_{\boldsymbol{\mu}} = \sum_{i=1}^{p} \boldsymbol{\mu}_i K_i$ for some $\boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\mu} \geq 0$. It is assumed that the kernels are uniformly bounded i.e. for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $i = 1 \ldots p$, we have $K_i(\mathbf{x}_1, \mathbf{x}_2) \leq \kappa_i^2$ for some $\kappa_i > 0$. Let $\boldsymbol{\kappa} = \left(\kappa_1^2, \ldots, \kappa_p^2\right) \in \mathbb{R}^p$. Note that $\boldsymbol{\kappa} \geq 0$. Also note that for any $\boldsymbol{\mu}$ and any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, we have $K_{\boldsymbol{\mu}}(\mathbf{x}_1, \mathbf{x}_2) \leq \langle \boldsymbol{\mu}, \boldsymbol{\kappa} \rangle$.

The notion of suitability used in [6] is that of *kernel-goodness* first proposed in [1] for classification tasks. For sake of simplicity, we shall henceforth consider only binary classification tasks, the extension to multi-class classification tasks being straightforward. We present below the notion of goodness used in [6]. For any binary classification task over a domain $\mathcal{X}$ characterized by a distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$, a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with associated Reproducing Kernel Hilbert Space $\mathcal{H}_K$ and feature map $\Phi_K : \mathcal{X} \to \mathcal{H}_K$ is said to be (ϵ, γ)-*kernel good* if there exists a unit norm vector $\mathbf{w} \in \mathcal{H}_K$ such that $\|\mathbf{w}\|_{\mathcal{H}_K} = 1$ and the following holds

$$\mathop{\mathbb{E}}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\!\!\left[\left[1 - \frac{y \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle}{\gamma} \right]_+ \right]\!\!\right] \leq \epsilon$$

2 Learning a *Good* Kernel

The key idea behind [6] is to try and learn a positive linear combination of kernels that is good according to the notion presented above. We define the risk functional $\mathcal{R}(\cdot) : \mathbb{R}^p \mapsto \mathbb{R}^+$ as follows:

$$\mathcal{R}(\boldsymbol{\mu}) := \mathop{\mathbb{E}}_{(\mathbf{x},y),(\mathbf{x}',y')\sim\mathcal{D}\times\mathcal{D}} \left[\!\left[1 - yy' K_{\boldsymbol{\mu}}(\mathbf{x},\mathbf{x}')\right]_+\right]$$

A combination $\boldsymbol{\mu}$ will be said to be *ϵ-combination good* if $\mathcal{R}(\boldsymbol{\mu}) \leq \epsilon$. The quantity $\mathcal{R}(\boldsymbol{\mu})$ is of interest since an application of Jensen's inequality (see [6, Lemma 3.2]) shows us that for any $\boldsymbol{\mu} \geq 0$ that is ϵ-combination good, the kernel $K_{\boldsymbol{\mu}}$ is $\left(\epsilon, \frac{1}{\langle\boldsymbol{\mu},\boldsymbol{\kappa}\rangle}\right)$-kernel good. Furthermore, one can show, using standard results on capacity of linear function classes (see for example [2, Theorem 21]), that an (ϵ,γ)-good kernel can be used to learn, with confidence $1-\delta$, a classifier with expected misclassification rate at most $\epsilon + \epsilon_1$ by using at most $\mathcal{O}\left(\frac{\kappa^4}{\epsilon_1^2\gamma^2}\log\frac{1}{\delta}\right)$ labeled samples.

In order to cast this learning problem more cleanly, [6] proposes the construction of a *K-space* using the following feature map

$$z : (\mathbf{x},\mathbf{x}') \mapsto \left(K_1(\mathbf{x},\mathbf{x}'),\ldots,K_p(\mathbf{x},\mathbf{x}')\right) \in \mathbb{R}^p$$

This allows us to write, for any $\boldsymbol{\mu} \in \mathbb{R}^p$, $K_{\boldsymbol{\mu}}(\mathbf{x},\mathbf{x}') = \langle\boldsymbol{\mu}, z(\mathbf{x},\mathbf{x}')\rangle$. Given n labeled training points $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)$, define the empirical risk functional $\hat{\mathcal{R}}(\cdot) : \mathbb{R}^p \mapsto \mathbb{R}^+$ as follows[1]:

$$\hat{\mathcal{R}}(\boldsymbol{\mu}) := \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} [1 - y_iy_j\langle\boldsymbol{\mu}, z(\mathbf{x}_i,\mathbf{x}_j)\rangle]_+$$

[6] poses the learning problem as the following optimization problem:

$$\min_{\boldsymbol{\mu}\geq 0} \frac{\lambda}{2}\|\boldsymbol{\mu}\|_2^2 + \hat{\mathcal{R}}(\boldsymbol{\mu})$$

3 Generalization Guarantees for a Learned Kernel Combination

Our generalization guarantee shall proceed in two steps. We shall assume that we have with us a training set $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)$ using which we are able to determine a combination vector $\hat{\boldsymbol{\mu}}$ such that $\hat{\mathcal{R}}(\hat{\boldsymbol{\mu}}) \leq \hat{\epsilon}$.

1. We shall first prove that, with high probability over the choice of the training points, the learned combination vector $\hat{\boldsymbol{\mu}}$ will give us a kernel $K_{\hat{\boldsymbol{\mu}}}$ that is $\left(\hat{\epsilon} + \epsilon_1, \frac{1}{\langle\hat{\boldsymbol{\mu}},\boldsymbol{\kappa}\rangle}\right)$-kernel good where $\epsilon_1 > 0$ is a quantity that can be made arbitrarily small.

---

[1]We note that [6] includes the terms $[1 - \langle\boldsymbol{\mu}, z(\mathbf{x}_i,\mathbf{x}_i)\rangle]_+$ into the empirical risk as well. This does not change the asymptotics of our analysis except for causing a bit of notational annoyance. In order to account for this term, the true risk functional will have to include an additional term $\mathcal{R}_{\text{add}}(\boldsymbol{\mu}) := \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\!\left[1 - K_{\boldsymbol{\mu}}(\mathbf{x},\mathbf{x})\right]_+\right]$. This will add a negligible term to the uniform convergence bound because we will have to consider the convergence of the term $\hat{\mathcal{R}}_{\text{add}}(\boldsymbol{\mu}) := \frac{2}{n(n+1)}\sum_{1\leq i\leq n}[1 - \langle\boldsymbol{\mu}, z(\mathbf{x}_i,\mathbf{x}_i)\rangle]_+$ to $\mathcal{R}_{\text{add}}$. However, from thereon, the analysis will remain unaffected since $\mathcal{R}_{\text{add}}(\boldsymbol{\mu}) \geq 0$ so a combination $\boldsymbol{\mu}$ having true risk $\mathcal{R}(\boldsymbol{\mu}) + \mathcal{R}_{\text{add}}(\boldsymbol{\mu}) \leq \epsilon$ will still give a kernel $K_{\boldsymbol{\mu}}$ that is $\left(\epsilon, \frac{1}{\langle\boldsymbol{\mu},\boldsymbol{\kappa}\rangle}\right)$-kernel good.

2. We shall then prove that given that there exists a good combination of kernels in the K-space, with very high probability $\hat{\epsilon}$ will be very small. This we will prove by showing a converse of the inequality proved in the first step. This will allow us to give oracle inequalities for the kernel goodness of the learned combination.

3.1 Step 1

In this step, we prove a uniform convergence guarantee for the learning problem at hand. Using standard proof techniques, we shall reduce the problem of uniform convergence to that of estimating the capacity of a certain function class. The notion of capacity we shall use is the Rademacher complexity which we shall bound using the heavy hammer of strong convexity based bounds from [5]. We note that the proof progression used in this step is fairly routine within the empirical process community and has been used to give generalization proofs for other problems as well (see for example [3, 4]).

First of all we note that due to the optimization process we have[2]

$$\frac{\lambda}{2} \|\hat{\boldsymbol{\mu}}\|_2^2 \leq \frac{\lambda}{2} \|\hat{\boldsymbol{\mu}}\|_2^2 + \hat{\mathcal{R}}(\hat{\boldsymbol{\mu}}) \leq \frac{\lambda}{2} \|\mathbf{0}\|_2^2 + \hat{\mathcal{R}}(\mathbf{0}) = 1$$

which implies that we need only concern ourselves with combination vectors inside the L_2 ball of radius $r_\lambda = \sqrt{\frac{2}{\lambda}}$.

$$\mathcal{B}_2\left(r_\lambda\right) := \{\boldsymbol{\mu} \in \mathbb{R}^p : \|\boldsymbol{\mu}\|_2 \leq r_\lambda\}$$

For notational simplicity, we denote $\mathbf{z} = (\mathbf{x}, y)$ as a training sample. For any training set $\mathbf{z}_1, \ldots, \mathbf{z}_n$ where $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ and for any $\boldsymbol{\mu} \in \mathbb{R}^p$, we write $\ell(\boldsymbol{\mu}, \mathbf{z}_i, \mathbf{z}_j) := [1 - y_i y_j \langle \boldsymbol{\mu}, z(\mathbf{x}_i, \mathbf{x}_j) \rangle]_+$. We assume, yet again for the sake of notational simplicity, that we obtain at all times, an even number of training samples i.e. n is even. For a ghost sample $\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n$ then, we can write

$$\mathbb{E}\left[\!\!\left[\frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \ell(\boldsymbol{\mu}, \tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j) \right]\!\!\right] = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathbb{E}\left[\!\!\left[\ell(\boldsymbol{\mu}, \tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j) \right]\!\!\right]$$

$$= \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathcal{R}(\hat{\boldsymbol{\mu}}) = \mathcal{R}(\hat{\boldsymbol{\mu}})$$

Thus we can write

$$\mathcal{R}(\hat{\boldsymbol{\mu}}) - \hat{\mathcal{R}}(\hat{\boldsymbol{\mu}}) = \mathbb{E}\left[\!\!\left[\frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \ell(\boldsymbol{\mu}, \tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j) \right]\!\!\right] - \hat{\mathcal{R}}(\hat{\boldsymbol{\mu}})$$

$$\leq \sup_{\boldsymbol{\mu} \in \mathcal{B}_2(r_\lambda)} \left\{ \mathbb{E}\left[\!\!\left[\frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \ell(\boldsymbol{\mu}, \tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j) \right]\!\!\right] - \hat{\mathcal{R}}(\boldsymbol{\mu}) \right\}$$

Let

$$g\left(\mathbf{z}_1, \ldots, \mathbf{z}_n\right) = \frac{2}{n(n-1)} \sup_{\boldsymbol{\mu} \in \mathcal{B}_2(r_\lambda)} \left\{ \mathbb{E}\left[\!\!\left[\sum_{1 \leq i < j \leq n} \ell(\boldsymbol{\mu}, \tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j) \right]\!\!\right] - \sum_{1 \leq i < j \leq n} \ell(\boldsymbol{\mu}, \mathbf{z}_i, \mathbf{z}_j) \right\}$$

---

[2]Any tolerance ϵ_{opt} incurred by the optimizer can easily be incorporated into the bounds. However, we do not do so for sake of clarity.

For any $\boldsymbol{\mu} \in \mathcal{B}_2(r_\lambda)$ and any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, we have $K_{\boldsymbol{\mu}}(\mathbf{x}_1, \mathbf{x}_2) \leq \langle \boldsymbol{\mu}, \boldsymbol{\kappa} \rangle \leq r_\lambda \|\boldsymbol{\kappa}\|_2$. Using this, it is not difficult to see that the expression $g(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ can be perturbed by at most $\frac{2}{n}(1 + r_\lambda \|\boldsymbol{\kappa}\|_2)$ by the change of a single true training sample $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ (see [6, Theorem 3.4] for the calculations). Applying McDiarmid's inequality to this expression, we get with probability at least $1 - \delta$,

$$\mathcal{R}(\hat{\boldsymbol{\mu}}) - \hat{\mathcal{R}}(\hat{\boldsymbol{\mu}}) \leq \mathbb{E}\left[\!\left[g(\mathbf{z}_1, \ldots, \mathbf{z}_n) \right]\!\right] + (1 + r_\lambda \|\boldsymbol{\kappa}\|_2)\sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

We now estimate the the expectation term on the right hand side.

$$
\begin{aligned}
\mathbb{E}\left[\!\left[g(\mathbf{z}_1, \ldots, \mathbf{z}_n) \right]\!\right] &= \frac{2}{n(n-1)} \mathbb{E}\left[\!\left[\sup_{\boldsymbol{\mu} \in \mathcal{B}_2(r_\lambda)} \left\{ \mathbb{E}\left[\!\left[\sum_{1 \leq i < j \leq n} \ell(\boldsymbol{\mu}, \tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j) \right]\!\right] - \sum_{1 \leq i < j \leq n} \ell(\boldsymbol{\mu}, \mathbf{z}_i, \mathbf{z}_j) \right\} \right]\!\right] \\
&\leq \frac{2}{n(n-1)} \mathbb{E}\left[\!\left[\sup_{\boldsymbol{\mu} \in \mathcal{B}_2(r_\lambda)} \left\{ \sum_{1 \leq i < j \leq n} \ell(\boldsymbol{\mu}, \tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j) - \sum_{1 \leq i < j \leq n} \ell(\boldsymbol{\mu}, \mathbf{z}_i, \mathbf{z}_j) \right\} \right]\!\right]
\end{aligned}
$$

We now invoke a powerful alternate representation for U-statistics to simplify the above expression. This method can be found in [4] that itself attributes this method to [8]. This, along with the Hoeffding decomposition, are two of the most powerful techniques to deal with "coupled" random variables as we have in this situation.

Theorem 1 ([4], Lemma A.1). *For any set of real valued functions $q_\tau : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ indexed by $\tau \in T$, if $X_1, \ldots, X_n$ are i.i.d. random variables then we have*

$$\mathbb{E}\left[\!\left[\sup_{\tau \in T} \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} q_\tau(X_i, X_j) \right]\!\right] \leq \mathbb{E}\left[\!\left[\sup_{\tau \in T} \frac{2}{n} \sum_{i=1}^{n/2} q_\tau(X_i, X_{n/2+i}) \right]\!\right]$$

Applying this decoupling result to the random variables $X_i = (\tilde{\mathbf{z}}_i, \mathbf{z}_i)$, the index set $\mathcal{B}_2(r_\lambda)$ and functions $q_\tau(X_i, X_j) = \ell(\boldsymbol{\mu}, \tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j) - \ell(\boldsymbol{\mu}, \mathbf{z}_i, \mathbf{z}_j)$ we get

$$
\begin{aligned}
\mathbb{E}\left[\!\left[g(\mathbf{z}_1, \ldots, \mathbf{z}_n) \right]\!\right] &\leq \frac{2}{n} \mathbb{E}\left[\!\left[\sup_{\boldsymbol{\mu} \in \mathcal{B}_2(r_\lambda)} \left\{ \sum_{i=1}^{n/2} \ell(\boldsymbol{\mu}, \tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_{n/2+i}) - \ell(\boldsymbol{\mu}, \mathbf{z}_i, \mathbf{z}_{n/2+i}) \right\} \right]\!\right] \\
&= \frac{2}{n} \mathbb{E}\left[\!\left[\sup_{\boldsymbol{\mu} \in \mathcal{B}_2(r_\lambda)} \left\{ \sum_{i=1}^{n/2} \epsilon_i \left(\ell(\boldsymbol{\mu}, \tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_{n/2+i}) - \ell(\boldsymbol{\mu}, \mathbf{z}_i, \mathbf{z}_{n/2+i}) \right) \right\} \right]\!\right] \\
&\leq \frac{4}{n} \mathbb{E}\left[\!\left[\sup_{\boldsymbol{\mu} \in \mathcal{B}_2(r_\lambda)} \left\{ \sum_{i=1}^{n/2} \epsilon_i \ell(\boldsymbol{\mu}, \mathbf{z}_i, \mathbf{z}_{n/2+i}) \right\} \right]\!\right] \\
&= \frac{4}{n} \mathbb{E}\left[\!\left[\sup_{\boldsymbol{\mu} \in \mathcal{B}_2(r_\lambda)} \left\{ \sum_{i=1}^{n/2} \epsilon_i \left[1 - y_i y_{n/2+i} \langle \boldsymbol{\mu}, z(\mathbf{x}_i, \mathbf{x}_{n/2+i}) \rangle \right]_+ \right\} \right]\!\right] \\
&\leq \frac{4}{n} \mathbb{E}\left[\!\left[\sup_{\boldsymbol{\mu} \in \mathcal{B}_2(r_\lambda)} \left\{ \sum_{i=1}^{n/2} \epsilon_i \langle \boldsymbol{\mu}, z(\mathbf{x}_i, \mathbf{x}_{n/2+i}) \rangle \right\} \right]\!\right] = 2\mathcal{R}_{n/2}(\mathcal{B}_2(r_\lambda))
\end{aligned}
$$

4

where in the second step, we performed symmetrization on the decoupled expression by introducing Rademacher random variables $\epsilon_i, i = 1, \ldots, n/2$. In the fifth step we have applied the contraction inequality stated in Theorem 2 below on the 1-Lipschitz function $\phi_i : x \mapsto = [1 - a_i x]_+$ where $a_i = y_i y_{n/2+i}$. We have exploited the fact that Theorem 2 actually proves the contraction inequality for the empirical Rademacher averages which allows us to treat a_i as constants dependent only on i.

Theorem 2. *Let $\mathcal{H}$ be a set of bounded real valued functions from some domain $\mathcal{X}$ and let $x_1, \ldots, x_n$ be arbitrary elements from $\mathcal{X}$. Furthermore, let $\phi_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n$ be L-Lipschitz functions. Then we have*

$$\mathbb{E}\left[\!\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \phi_i(h(x_i))\right]\!\right] \leq L \mathbb{E}\left[\!\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(x_i)\right]\!\right]$$

Proof. Ledoux and Talagrand (see [7, Theorem 4.12]) prove the same result but for wrapper functions that satisfy $\phi_i(0) = 0$ for all i. To get the result, simply apply the result to the functions $\tilde{\phi}_i : x \mapsto \phi_i(x) - \phi_i(0)$ to get

$$\mathbb{E}\left[\!\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \phi_i(h(x_i))\right]\!\right] \leq \mathbb{E}\left[\!\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \tilde{\phi}_i(h(x_i))\right]\!\right] + \mathbb{E}\left[\!\left[\frac{1}{n} \sum_{i=1}^n \epsilon_i \phi_i(0)\right]\!\right] \leq L \mathbb{E}\left[\!\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(x_i)\right]\!\right]$$

where we apply [7, Theorem 4.12] to the first term and the second term vanishes by linearity of expectation. $\square$

The concluding term in the last chain of inequalities gives us the Rademacher complexity of the hypothesis class $\mathcal{B}_2(r_\lambda)$. At this point we introduce the following result on Rademacher complexities of regularized linear predictor classes

Theorem 3 ([5], Theorem 1). *Let $\mathcal{W}$ be a closed convex set and let $F : \mathcal{W} \to \mathbb{R}$ be λ-strongly convex w.r.t. $\|\cdot\|_*$. Assume $\mathcal{W} \subseteq \{\mathbf{w} : F(\mathbf{w}) \leq W_*^2\}$. Furthermore, let $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\| \leq X\}$ and $\mathcal{F}_{\mathcal{W}} := \{\mathbf{w} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \mathbf{w} \in \mathcal{W}, \mathbf{x} \in \mathcal{X}\}$. Then, we have*

$$\hat{\mathcal{R}}_n(\mathcal{F}_{\mathcal{W}}) \leq X W_* \sqrt{\frac{2}{\lambda n}}$$

Although [5] make their claim for the normal Rademacher average but their proof actually gives bounds for the empirical Rademacher averages. Since our hypothesis class is L_2 regularized, we can apply Theorem 3 to the L_2/L_2 case with $F(\boldsymbol{\mu}) = \|\boldsymbol{\mu}\|_2^2$ as the regularizer. Since we have $\sup_{\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}} \|z(\mathbf{x}_1, \mathbf{x}_2)\|_2 \leq \|\boldsymbol{\kappa}\|_2$, we get

$$\mathcal{R}_{n/2}(\mathcal{B}_2(r_\lambda)) \leq r_\lambda \|\boldsymbol{\kappa}\|_2 \sqrt{\frac{2}{n}} = 2 \|\boldsymbol{\kappa}\|_2 \sqrt{\frac{1}{\lambda n}}$$

We have thus proved the following result

Theorem 4. *With probability at least $1 - \delta$ over the choice of training samples, the minimizer $\hat{\boldsymbol{\mu}}$ of the expression*

$$\min_{\boldsymbol{\mu} \geq 0} \frac{\lambda}{2} \|\boldsymbol{\mu}\|_2^2 + \hat{\mathcal{R}}(\boldsymbol{\mu})$$

satisfies the following

$$\mathcal{R}(\boldsymbol{\mu}) \leq \hat{\mathcal{R}}(\boldsymbol{\mu}) + 4 \left\|\boldsymbol{\kappa}\right\|_2 \sqrt{\frac{1}{\lambda n}} + \left(1 + \left\|\boldsymbol{\kappa}\right\|_2 \sqrt{\frac{2}{\lambda}}\right) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}} \leq \hat{\mathcal{R}}(\boldsymbol{\mu}) + 6 \left\|\boldsymbol{\kappa}\right\|_2 \sqrt{\frac{\log \frac{1}{\delta}}{\lambda n}}$$

Since $\hat{\boldsymbol{\mu}} \in \mathcal{B}_2(r_\lambda)$, we have $\langle \hat{\boldsymbol{\mu}}, \boldsymbol{\kappa} \rangle \leq \sqrt{\frac{2}{\lambda}} \left\|\boldsymbol{\kappa}\right\|_2$. This implies that the kernel $K_{\hat{\boldsymbol{\mu}}}$ is at least $\left(\hat{\epsilon} + \epsilon_1, \frac{1}{\|\boldsymbol{\kappa}\|_2} \sqrt{\frac{\lambda}{2}}\right)$-kernel good where $\hat{\epsilon} = \hat{\mathcal{R}}(\boldsymbol{\mu})$ and $\epsilon_1 \leq 6 \left\|\boldsymbol{\kappa}\right\|_2 \sqrt{\frac{\log \frac{1}{\delta}}{\lambda n}}$. In particular, if all the p kernels share a common bound i.e. $\kappa_i \leq \kappa$ for all i, then $\left\|\boldsymbol{\kappa}\right\|_2 \leq \kappa^2 \sqrt{p}$ and we can show the kernel $K_{\hat{\boldsymbol{\mu}}}$ to be $\left(\hat{\epsilon} + 6\kappa^2 \sqrt{\frac{p \log \frac{1}{\delta}}{\lambda n}}, \frac{1}{\kappa^2} \sqrt{\frac{\lambda}{2p}}\right)$-kernel good.

3.2 Step 2

Just as we analyzed the excess risk expression $\mathcal{R}(\boldsymbol{\mu}) - \hat{\mathcal{R}}(\boldsymbol{\mu})$ uniformly over vectors the ball $\mathcal{B}_2(r_\lambda)$, we can similarly analyze the expression $\hat{\mathcal{R}}(\boldsymbol{\mu}) - \mathcal{R}(\boldsymbol{\mu})$ uniformly over any (fixed) ball $\mathcal{B}_2(r)$ to get the following result.

Theorem 5. *Let $r > 0$ be some fixed radius, then with probability at least $1 - \delta$ over the choice of training samples, all combination vectors $\boldsymbol{\mu} \in \mathcal{B}_2(r)$ satisfy*

$$\hat{\mathcal{R}}(\boldsymbol{\mu}) \leq \mathcal{R}(\boldsymbol{\mu}) + 2r \left\|\boldsymbol{\kappa}\right\|_2 \sqrt{\frac{2}{n}} + (1 + r \left\|\boldsymbol{\kappa}\right\|_2) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

This allows us to give the following oracle inequality:

Theorem 6. *Suppose as an oracle assumption we assume that there exists a good combination vector $\boldsymbol{\mu}_o$ that is ϵ_o-combination good, then we can output with probability at least $1 - \delta$, for any $\epsilon_1 > 0$ using $n = \Omega\left(\frac{\|\boldsymbol{\mu}_o\|_2^2}{\epsilon_1^3}\right)$ training samples, a combination vector such that the corresponding kernel that is $\left(\epsilon_o + \epsilon_1, \frac{1}{\|\boldsymbol{\kappa}\|_2 \|\boldsymbol{\mu}_o\|_2} \sqrt{\frac{\epsilon_1}{3}}\right)$-kernel good.*

Proof. Using Theorem 5 we have with probability at least $1 - \delta$,

$$\hat{\mathcal{R}}(\boldsymbol{\mu}_o) \leq \epsilon_o + 2 \left\|\boldsymbol{\mu}_o\right\|_2 \left\|\boldsymbol{\kappa}\right\|_2 \sqrt{\frac{2}{n}} + (1 + \left\|\boldsymbol{\mu}_o\right\|_2 \left\|\boldsymbol{\kappa}\right\|_2) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}} \leq \epsilon_o + 6 \left\|\boldsymbol{\mu}_o\right\|_2 \left\|\boldsymbol{\kappa}\right\|_2 \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

Since $\hat{\boldsymbol{\mu}}$ is the minimizer of the regularized empirical risk, we have

$$\frac{\lambda}{2} \left\|\hat{\boldsymbol{\mu}}\right\|_2^2 + \hat{\mathcal{R}}(\hat{\boldsymbol{\mu}}) \leq \frac{\lambda}{2} \left\|\boldsymbol{\mu}_o\right\|_2^2 + \hat{\mathcal{R}}(\boldsymbol{\mu}_o) \leq \frac{\lambda}{2} \left\|\boldsymbol{\mu}_o\right\|_2^2 + \epsilon_o + 6 \left\|\boldsymbol{\mu}_o\right\|_2 \left\|\boldsymbol{\kappa}\right\|_2 \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

which gives us, since $\left\|\hat{\boldsymbol{\mu}}\right\|_2 \geq 0$,

$$\hat{\mathcal{R}}(\hat{\boldsymbol{\mu}}) \leq \epsilon_o + \frac{\lambda}{2} \left\|\boldsymbol{\mu}_o\right\|_2^2 + 6 \left\|\boldsymbol{\mu}_o\right\|_2 \left\|\boldsymbol{\kappa}\right\|_2 \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

6

Applying Theorem 4 we get with probability at least $1 - 2\delta$,

$$\mathcal{R}(\hat{\boldsymbol{\mu}}) \le \epsilon_o + \frac{\lambda}{2} \|\boldsymbol{\mu}_o\|_2^2 + 6 \|\boldsymbol{\mu}_o\|_2 \|\boldsymbol{\kappa}\|_2 \sqrt{\frac{2 \log \frac{1}{\delta}}{n}} + 6 \|\boldsymbol{\kappa}\|_2 \sqrt{\frac{\log \frac{1}{\delta}}{\lambda n}}$$

For any $0 < \epsilon_1 < 3/4$, setting $\lambda = \frac{2\epsilon_1}{3\|\boldsymbol{\mu}_o\|_2^2}$ and requiring $n \ge \frac{500}{\epsilon_1^3} \|\boldsymbol{\mu}_o\|_2^2 \|\boldsymbol{\kappa}\|_2^2 \log \frac{1}{\delta}$ so that all three terms in the above expression are less than $\epsilon_1/3$ gives us the result (for values of ϵ_1 larger than $3/4$, $n \ge \frac{650}{\epsilon_1^2} \|\boldsymbol{\mu}_o\|_2^2 \|\boldsymbol{\kappa}\|_2^2 \log \frac{1}{\delta}$ suffices). $\qquad\square$

Such oracle inequalities are very desirable since they tell us that we would be able to give a performance that is competitive against any fixed kernel in foresight. If we set λ to an oracle oblivious value such as $\lambda = \sqrt[3]{\frac{1}{n}}$ then although we get an inferior claim with respect to the kernel-goodness, we are able to make that claim in hindsight as well.

4 Learning Sparse Kernel Combinations

Since the complexity of the evaluating the kernel $K_{\boldsymbol{\mu}}$ goes up roughly as $\|\boldsymbol{\mu}\|_0$, it is desirable to learn sparse combinations. This can be done by changing the learning formulation slightly to the following:

$$\min_{\boldsymbol{\mu} \ge 0} \frac{\lambda}{2} \|\boldsymbol{\mu}\|_1 + \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} [1 - y_i y_j \langle \boldsymbol{\mu}, z(\mathbf{x}_i, \mathbf{x}_j) \rangle]_+$$

The above learning algorithm can also shown to admit generalization guarantees. For sake of brevity we only give below the main points where the analysis differs from the L_2 regularized case. First of all, we would be able to show that the regularized empirical risk minimizer $\hat{\boldsymbol{\mu}}$ would lie in the L_1 ball $\mathcal{B}_1(s_\lambda) := \{\boldsymbol{\mu} \in \mathbb{R}^p : \|\boldsymbol{\mu}\|_1 \le s_\lambda\}$ where $s_\lambda = \frac{2}{\lambda}$.

Due to this the perturbations to the expression $g(\mathbf{z}_1, \dots, \mathbf{z}_n)$ would be limited by $\frac{2}{n}(1 + s_\lambda \|\boldsymbol{\kappa}\|_\infty)$. While applying Theorem 3, we would instead consider the regularizer $F(\boldsymbol{\mu}) = \|\boldsymbol{\mu}\|_q^2$ for $q = \frac{\log p}{\log p - 1}$ which is $\left(\frac{1}{\log p}\right)$-strongly convex with respect to the norm $\|\cdot\|_1$. This would allow us to bound the Rademacher complexity of the hypothesis class $\mathcal{B}_1(s_\lambda)$ as

$$\mathcal{R}_{n/2}(\mathcal{B}_1(s_\lambda)) \le s_\lambda \|\boldsymbol{\kappa}\|_\infty \sqrt{\frac{2 \log p}{n}} = \frac{2 \|\boldsymbol{\kappa}\|_\infty}{\lambda} \sqrt{\frac{2 \log p}{n}}$$

This allows us to make the following claim:

Theorem 7. *With probability at least $1 - \delta$ over the choice of training samples, the minimizer $\hat{\boldsymbol{\mu}}$ of the expression*

$$\min_{\boldsymbol{\mu} \ge 0} \frac{\lambda}{2} \|\boldsymbol{\mu}\|_1 + \hat{\mathcal{R}}(\boldsymbol{\mu})$$

satisfies the following

$$\mathcal{R}(\boldsymbol{\mu}) \le \hat{\mathcal{R}}(\boldsymbol{\mu}) + \frac{4 \|\boldsymbol{\kappa}\|_\infty}{\lambda} \sqrt{\frac{2 \log p}{n}} + \left(1 + \frac{2 \|\boldsymbol{\kappa}\|_\infty}{\lambda}\right) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}} \le \hat{\mathcal{R}}(\boldsymbol{\mu}) + \frac{6 \|\boldsymbol{\kappa}\|_\infty}{\lambda \sqrt{n}} \left(\sqrt{\log p} + \sqrt{\log 1/\delta}\right)$$

Since $\hat{\boldsymbol{\mu}} \in \mathcal{B}_1(s_\lambda)$, we have $\langle \hat{\boldsymbol{\mu}}, \boldsymbol{\kappa} \rangle \leq \frac{2\|\boldsymbol{\kappa}\|_\infty}{\lambda}$. This implies that the kernel $K_{\hat{\boldsymbol{\mu}}}$ is $\left(\hat{\epsilon} + \epsilon_1, \frac{\lambda}{2\|\boldsymbol{\kappa}\|_\infty} \right)$-kernel good where $\hat{\epsilon} = \hat{\mathcal{R}}(\boldsymbol{\mu})$ and $\epsilon_1 \leq \frac{6\|\boldsymbol{\kappa}\|_\infty}{\lambda\sqrt{n}} \left(\sqrt{\log p} + \sqrt{\log 1/\delta} \right)$. In particular, if all the p kernels share a common bound i.e. $\kappa_i \leq \kappa$ for all i, then $\|\boldsymbol{\kappa}\|_\infty \leq \kappa^2$ and we can show the kernel $K_{\hat{\boldsymbol{\mu}}}$ to be $\left(\hat{\epsilon} + \frac{6\kappa^2}{\lambda\sqrt{n}} \left(\sqrt{\log p} + \sqrt{\log 1/\delta} \right), \frac{\lambda}{2\kappa^2} \right)$-kernel good.

Note that this result has a much better dependence on p that the result for L_2 regularized learning where we were only able to show that the kernel $K_{\hat{\boldsymbol{\mu}}}$ was $\left(\hat{\epsilon} + 6\kappa^2 \sqrt{\frac{p \log \frac{1}{\delta}}{\lambda n}}, \frac{1}{\kappa^2} \sqrt{\frac{\lambda}{2p}} \right)$-kernel good.

We can also show the following version of Theorem 5 to be true

Theorem 8. *Let $s > 0$ be some fixed radius, then with probability at least $1 - \delta$ over the choice of training samples, all combination vectors $\boldsymbol{\mu} \in \mathcal{B}_1(s)$ satisfy*

$$\hat{\mathcal{R}}(\boldsymbol{\mu}) \leq \mathcal{R}(\boldsymbol{\mu}) + 2s \|\boldsymbol{\kappa}\|_\infty \sqrt{\frac{2 \log p}{n}} + (1 + s \|\boldsymbol{\kappa}\|_\infty) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

Using this, and going as before, we are also able to guarantee the following oracle inequality similar to Theorem 6

Theorem 9. *Suppose as an oracle assumption we assume that there exists a good combination vector $\boldsymbol{\mu}_o$ that is ϵ_o-combination good, then we can output with probability at least $1 - \delta$, for any $\epsilon_1 > 0$ using $n = \Omega \left(\frac{\|\boldsymbol{\mu}_o\|_1}{\epsilon_1^2} \right)$ training samples, a combination vector such that the corresponding kernel that is $\left(\epsilon_o + \epsilon_1, \frac{\epsilon_1}{3\|\boldsymbol{\kappa}\|_\infty \|\boldsymbol{\mu}_o\|_1} \right)$-kernel good.*

Proof. Following the chain of inequalities given by Theorems 7 and 8 and using optimality of the regularized empirical risk minimizer $\hat{\boldsymbol{\mu}}$, we get, with probability at least $1 - 2\delta$,

$$\mathcal{R}(\hat{\boldsymbol{\mu}}) \leq \epsilon_o + \frac{\lambda}{2} \|\boldsymbol{\mu}_o\|_1 + \frac{6 \|\boldsymbol{\kappa}\|_\infty \left(\sqrt{\log p} + \sqrt{\log 1/\delta} \right)}{\sqrt{n}} \left(\|\boldsymbol{\mu}_o\|_1 + \frac{1}{\lambda} \right)$$

Setting $\lambda = \frac{2\epsilon_1}{3\|\boldsymbol{\mu}_o\|_1}$ and requiring $n \geq \frac{135\|\boldsymbol{\mu}_o\|_1 \|\boldsymbol{\kappa}\|_\infty \left(\sqrt{\log p} + \sqrt{\log 1/\delta} \right)}{\epsilon_1^2}$ finishes the proof. $\qquad \square$

5   Discussion on the Nature of Guarantees

The guarantees given above, both for the sparse as well as the non-sparse kernel learning cases are slightly unsatisfactory in the sense they assume combination goodness to ensure kernel goodness. In other words they assume the existence of a combination that is ϵ-combination before guaranteeing that the output would be a kernel that is (ϵ', γ')-kernel good. Ideally, we should have used the promise of existence of a kernel that (ϵ, γ)-kernel good to ensure that a good kernel is output.

One way to prove such a result would be to show that if there exists a kernel combination that is (ϵ, γ)-kernel good, then there also exists some combination $\boldsymbol{\mu} \in \mathbb{R}^p$ that is ϵ'-combination good for some $\epsilon' > 0$. However, this is an unlikely result and the the aim of this section is to discuss this point. It turns out that the biggest hurdle that one faces in proving such a result is the form of

combination goodness chosen by [6]. The definition of combination goodness used in [6] is related to the notion of *similarity goodness* proposed in [1] except for the absence of a *weight function.*

More specifically, [1] consider a kernel $K_{\boldsymbol{\mu}}$ to be ϵ-similarity good if for some weight function $w : \mathcal{X} \to \mathbb{R}$ the following holds:

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\!\!\left[\left[1 - y \mathop{\mathbb{E}}_{(\mathbf{x}',y')\sim\mathcal{D}} \left[\!\!\left[y' w(\mathbf{x}') K_{\boldsymbol{\mu}}(\mathbf{x},\mathbf{x}') \right]\!\!\right] \right]_+ \right]\!\!\right] \le \epsilon$$

For ease of comparison, we have absorbed the margin parameter γ in the definition given in [1] into the weight function $w(\cdot)$. Note that if the notion of combination goodness had been defined using

$$\mathcal{R}(\boldsymbol{\mu}) := \mathop{\mathbb{E}}_{(\mathbf{x},y),(\mathbf{x}',y')\sim\mathcal{D}\times\mathcal{D}} \left[\!\!\left[\left[1 - yy' w(\mathbf{x}') K_{\boldsymbol{\mu}}(\mathbf{x},\mathbf{x}') \right]_+ \right]\!\!\right]$$

instead, then one could have used some form of inverse Jensen's inequality to convert similarity goodness into combination goodness. Since the presence of the weight function makes it possible for crisp conversions of kernel goodness into similarity goodness as was done in [9], this could have been one way to convert kernel goodness into combination goodness (i.e. via similarity goodness). However, due to the absence of such weight functions, it seems difficult to convert kernel goodness into combination goodness using the methods of [9].

Another reason to believe in the non-existence of such conversions from kernel to combination goodness is the form of the predictor in the RKHS. If one looks at the proof of Lemma 3.2 in [6] then one notices that the kernel goodness is proven with respect to the predictor $\mathbf{w} = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\!\!\left[y' \Phi_{\mathcal{H}_{K_{\boldsymbol{\mu}}}}(\mathbf{x}) \right]\!\!\right]$ where $\Phi_{\mathcal{H}_{K_{\boldsymbol{\mu}}}} : \mathcal{X} \mapsto \mathcal{H}_{K_{\boldsymbol{\mu}}}$ is the feature map corresponding to the kernel $K_{\boldsymbol{\mu}}$. This turns out to be very a restrictive form for the predictor. A kernel can be good due to the existence of *any* unit norm predictor in its RKHS. However the notion of combination goodness seems to prefer predictors that point from the mean of the images of the negative points to the mean of the images of the positive points in the RHKS. It was noted in [1] that such a notion of goodness is too strong ([1] actually call this the *strongly-good* notion of similarity goodness) and that there exist kernels that are very good with respect to the learning task at hand but the uniform vectors $\mathbf{w} = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\!\!\left[y' \Phi_{\mathcal{H}_{K_{\boldsymbol{\mu}}}}(\mathbf{x}) \right]\!\!\right]$ in their RKHSes perform poorly (see [1, Definition 2] and the discussion thereafter).

Thus it seems unlikely that the current proof technique can be extended to accept promises of kernel goodness. The technique seems inherently suited to accept combination goodness and output good kernels. It would be interesting to see whether the existing proofs can be modified or whether the algorithms can be modified to accommodate kernel goodness.

References

[1] Maria-Florina Balcan and Avrim Blum. On a Theory of Learning with Similarity Functions. In *International Conference on Machine Learning*, 2006.

[2] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[3] Qiong Cao, Zheng-Chu Guo, and Yiming Ying. Generalization Bounds for Metric and Similarity Learning, 2012. arXiv:1207.5437.

[4] Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *Annals of Statistics*, 36:844–874, 2008.

[5] Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization. In *Annual Conference on Neural Information Processing Systems*, 2008.

[6] Abhishek Kumar, Alexandru Niculescu-Mizil, Koray Kavukcuoglu, and Hal Daumé III. A Binary Classification Framework for Two-Stage Multiple Kernel Learning. In *International Conference on Machine Learning*, 2012.

[7] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 2002.

[8] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

[9] Nathan Srebro. How Good Is a Kernel When Used as a Similarity Measure? In *20th Annual Conference on Computational Learning Theory*, pages 323–335, 2007.