Optimality Implies Kernel Sum Classifiers are Statistically Efficient

Raphael A. Meyer [1] **Jean Honorio** [1]

Abstract

We propose a novel combination of optimization tools with learning theory bounds in order to analyze the sample complexity of *optimal* kernel sum classifiers. This contrasts the typical learning theoretic results which hold for all (potentially suboptimal) classifiers. Our work also justifies assumptions made in prior work on multiple kernel learning. As a byproduct of our analysis, we also provide a new form of Rademacher complexity for hypothesis classes containing only optimal classifiers.

1. Introduction

Classification is a fundamental task in machine learning (Shalev-Shwartz & Ben-David, 2014; Daumé III, 2012; Friedman et al., 2001). Kernel methods allow classifiers to learn powerful nonlinear relationships (Shawe-Taylor et al., 2004; Balcan et al., 2006). Optimization tools allow these methods to learn efficiently (Soentpiet et al., 1999). Under mild assumptions, kernels guarantee that learned models generalize well (Bartlett & Mendelson, 2002). However, the overall quality of these models still depends heavily on the choice of kernel. To compensate for this, prior work considers learning how to linearly combine a set of arbitrary kernels into a good data-dependent kernel (Sonnenburg et al., 2006; Gönen & Alpaydın, 2011; Bach et al., 2004).

It is known that if the learned linear combination of kernels is well behaved, then the kernel classifier generalizes well (Cortes et al., 2010; 2009a; Argyriou et al., 2005). We extend this body of work by proving that if our classifier is optimal, then the linear combination of kernels is well behaved. This optimality assumption is well justified because many common machine learning problems are solved using optimization algorithms. For instance, in this paper we consider binary classification with Kernel Support Vector Machines (SVM), which are computed by solving a quadratic programming problem. Specifically, we bound the sample complexity of kernel classifiers in two regimes. In the first, we are forced to classify using the sum of a set of kernels. In the second, we choose which kernels we include in our summation.

There exists substantial prior work considering learning kernels. From the computational perspective, several theoretically sound and experimentally efficient algorithms are known (Cortes et al., 2009b;a; Kivinen et al., 2004; Sinha & Duchi, 2016; Duvenaud et al., 2013). Much of this work relies on optimization tools such as quadratic programs (Chen et al., 2009), sometimes specifically considering Kernel SVM (Srebro & Ben-David, 2006). This motivates our focus on optimal classifiers for multiple kernel learning. The literature on sample complexity for these problems always assumes that the learned combination of kernels is well behaved (Cortes et al., 2009c; 2012; 2013; Srebro & Ben-David, 2006; Sinha & Duchi, 2016). That is, the prior work assumes that the weighted sum of kernel matrices $\tilde{\boldsymbol{K}}_\Sigma$ is paired with a vector $\boldsymbol{\alpha}_\Sigma$ such that $\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \leq C^2$ for some constant C. It is unclear how C depends on the structure or number of base kernels. Our work provides bounds that explains this relationship for optimal classifiers. Additionally, Rademacher complexity is typically used to control the generalization error over all possible (not necessarily optimal) estimators (Bartlett & Mendelson, 2002; Koltchinskii et al., 2002; Kakade et al., 2009). We differ from this approach by bounding the Rademacher complexity for only optimal estimators. We are not aware of any prior work that explores such bounds.

Contributions. Our results start with a core technical theorem, which is then applied to two novel hypothesis classes.

- We first show that the optimal solution to the Kernel SVM problem using the sum of m kernels is well behaved. That is, we consider the given kernel matrices $\tilde{\boldsymbol{K}}_1, \ldots, \tilde{\boldsymbol{K}}_m$ and the corresponding Dual Kernel SVM solution vectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m$, as well as the sum of these kernel matrices $\tilde{\boldsymbol{K}}_\Sigma$ and its Dual Kernel SVM solution vector $\boldsymbol{\alpha}_\Sigma$. Using Karush-Kuhn-Tucker (KKT) optimality conditions, we prove that

[1]Department of Computer Science, Purdue University, Indiana, USA. Correspondence to: Raphael A. Meyer <meyer219@purdue.edu>, Jean Honorio <jhonorio@purdue.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

$\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \leq 3m^{-0.58}B^2$ provided that all base kernels fulfill $\boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t \leq B^2$ for some constant B. We are not aware of any existing bounds of this kind, and we provide Rademacher complexity analysis to leverage this result. Note that the previous bounds for the Rademacher complexity in multiple kernel learning assumes that $\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma$ is bounded.

We provide Rademacher complexity bounds for two novel hypothesis classes. As opposed to traditional Rademacher bounds, our hypothesis classes only contain optimal classifiers. The traditional analysis when using a *single* kernel provides an empirical Rademacher complexity bound of $O(\frac{BR}{\sqrt{n}})$, where n is the number of samples and $k_t(\mathbf{x}_i, \mathbf{x}_i) \leq R^2$ bounds the radius of the samples in every feature space (Bartlett & Mendelson, 2002).

- **Kernel Sums:** In the first set, Kernel SVM is required to use the sum of all m kernels. We show that the empirical Rademacher complexity is bounded by $O(\frac{BR}{\sqrt{n}}m^{0.208})$.
- **Kernel Subsets:** In the second set, Kernel SVM is allowed to use the sum of any subset of the m kernels. The classical analysis in this setting would pay a multiplicative factor of 2^{m-1}. The approach we use instead only pays with a factor of $\sqrt{\ln(m)}$. We prove that the empirical Rademacher complexity is bounded by $O(\frac{BR\sqrt{\ln(m)}}{\sqrt{n}}m^{0.208})$.

Note that these Rademacher bounds compare naturally to the traditional single kernel bound. If we use a sum of m kernels instead of just one kernel, then we pay a multiplicative factor of $m^{0.208}$. If we use any subset of kernels, we only pay an extra factor of $\sqrt{\ln(m)}$. Thus, in this work, we show that optimization tools such as KKT conditions are useful in the analysis of statistical bounds. These optimization bounds are leveraged by learning theoretic tools such as Rademacher complexity, as seen in the second and third bullet points. Overall, we obtain new bounds with natural assumptions that connect the existing literature on optimization and learning theory in a novel fashion. Additionally, these bounds justify assumptions made in the existing literature.

2. Preliminaries

Let $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ denote a dataset of n i.i.d. samples from some distribution $\mathcal{D}$, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \{-1, 1\}$ for some $\mathcal{X}$. Let $\|\cdot\|_2$ denote the ℓ_2 vector norm and $\|\cdot\|_1$ denote the ℓ_1 vector norm. Let $[n] := \{1, \ldots, n\}$ for any natural number n.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote a kernel function. In this paper, we assume that all kernels fulfill $|k(\mathbf{x}, \mathbf{x})| < \infty$ for all $\mathbf{x} \in \mathcal{X}$. We consider being given a set of kernels $k_1, \ldots, k_m$. Let $k_\Sigma(\cdot,\cdot) := \sum_{t=1}^m k_t(\cdot,\cdot)$ denote the sum of the m kernels. The above notation will be useful when learning with kernel sums. Let $\mathcal{P} \subseteq [m]$. Then define $k_\mathcal{P}(\cdot,\cdot) := \sum_{t\in\mathcal{P}} k_t(\cdot,\cdot)$ as the sum of kernels as described by $\mathcal{P}$. The latter notation will be useful when learning kernel subsets.

Given a dataset $\mathcal{S}$ and a kernel k_t, we can build the corresponding kernel matrix $\boldsymbol{K}_t \in \mathbb{R}^{n\times n}$, where $[\boldsymbol{K}_t]_{i,j} := k_t(\mathbf{x}_i, \mathbf{x}_j)$. Further, we can build the *labeled kernel matrix* $\tilde{\boldsymbol{K}}_t$, defined elementwise as $[\tilde{\boldsymbol{K}}_t]_{i,j} := y_i y_j k_t(\mathbf{x}_i, \mathbf{x}_j)$. To simplify notation, all our results use labeled kernel matrices instead of standard kernel matrices.

2.1. Separable SVM

We now present optimal kernel classification, first in the separable case.

Definition 1 (Primal Kernel SVM)**.** *Given a dataset $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and a feature map $\boldsymbol{\phi} : \mathcal{X} \mapsto \mathbb{R}^d$, the Primal Kernel SVM problem is equivalent to the following optimization problem:*

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$
$$\text{s.t. } 1 - y_i \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_i) \leq 0 \ \ \forall i \in [n]$$

We will mainly look at the corresponding dual problem:

Definition 2 (Dual Kernel SVM)**.** *Given a dataset $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and a kernel function $k(\cdot,\cdot)$ with associated labeled kernel matrix $\tilde{\boldsymbol{K}}$, the Dual Kernel SVM problem is equivalent to the following optimization problem:*

$$\max_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 - \frac{1}{2}\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha}$$
$$\text{s.t. } \alpha_i \geq 0 \ \ \forall i \in [n]$$

Since the dual optimization problem is defined entirely in terms of $\tilde{\boldsymbol{K}}$, we can denote the optimal $\boldsymbol{\alpha}$ as a function of the labeled kernel matrix. We write this as $\boldsymbol{\alpha} = \text{DualSVM}(\tilde{\boldsymbol{K}})$.

Recall that Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient for optimality in convex optimization problems (Boyd & Vandenberghe, 2004). We can express the KKT conditions of the Primal Kernel SVM as follows:

Primal Feasibility:

$$1 - y_i \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_i) \leq 0 \ \ \forall i \in [n] \tag{1}$$

Stationarity:

$$\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \boldsymbol{\phi}(\mathbf{x}_i) \tag{2}$$

Dual Feasibility:

$$\alpha_i \geq 0 \ \ \forall i \in [n] \tag{3}$$

Complementary Slackness:

$$\alpha_i(1 - y_i \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_i)) = 0 \tag{4}$$

The above KKT conditions will be used with learning theoretic tools in order to provide novel generalization bounds.

2.2. Non-separable SVM

The primal and dual SVMs above assume that the given kernel is able to separate the data perfectly. Since this is not always the case, we also consider non-separable data using ℓ_2 slack variables:

Definition 3 (Primal Kernel SVM with Slack Variables)**.** *Given $C > 0$, a dataset $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, and a feature map $\boldsymbol{\phi} : \mathcal{X} \mapsto \mathbb{R}^d$, the Primal Kernel SVM problem is equivalent to the following optimization problem:*

$$\begin{aligned} \min_{\mathbf{w}} \ & \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{C}{2}\|\boldsymbol{\xi}\|_2^2 \\ \textit{s.t.} \ & 1 - y_i \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_i) \leq \xi_i \ \ \forall i \in [n] \\ & \xi_i \geq 0 \ \ \forall i \in [n] \end{aligned}$$

Definition 4 (Dual Kernel SVM with Slack Variables)**.** *Given $C > 0$, a dataset $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and a kernel function $k(\cdot, \cdot)$ with associated labeled kernel matrix $\tilde{\boldsymbol{K}}$, the Dual Kernel SVM problem is equivalent to the following optimization problem:*

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \boldsymbol{\xi}} \ & \|\boldsymbol{\alpha}\|_1 - \frac{1}{2}\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha} - \frac{1}{2}\|\boldsymbol{\xi}\|_2^2 \\ \textit{s.t.} \ & 0 \leq \alpha_i \leq C\xi_i \ \ \forall i \in [n] \end{aligned}$$

We denote the solution to the Dual SVM with Slack Variables using parameter C as $\boldsymbol{\alpha} = \text{DualSVM}_C(\tilde{\boldsymbol{K}})$.

2.3. Rademacher Complexity for Kernels

We use Rademacher complexity to bound the sample complexity of kernel methods. The empirical Rademacher complexity of a hypothesis class $\mathcal{F}$ with dataset $\mathcal{S}$ is defined as

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} \sigma_i h(\mathbf{x}_i)\right)\right]$$

where $\boldsymbol{\sigma} \in \{-1, +1\}^n$ is a vector of Rademacher variables. Bartlett and Mendelson introduced the analysis of sample complexity for kernel methods via Rademacher complexity when using one kernel (2002). Bartlett and Mendelson considered the following hypothesis class of representer theorem functions:

$$\mathcal{F} := \left\{\mathbf{x} \mapsto \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i), \mathbf{x}_i \in \mathcal{X}, \boldsymbol{\alpha}^\intercal \boldsymbol{K} \boldsymbol{\alpha} \leq B^2\right\} \tag{5}$$

Each element of $\mathcal{F}$ is defined in terms of a dataset $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and an $\boldsymbol{\alpha}$ vector. Bartlett and Mendelson showed that the probability of misclassification is bounded by the empirical risk of misclassification with a γ-Lipschitz loss plus a Rademacher term:

Theorem 1 (Bartlett & Mendelson 2002)**.** *Fix $n \geq 0$, $\gamma \in (0, 1)$, and $\delta \in (0, 1)$.*

Let $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be a dataset of n i.i.d. samples from $\mathcal{D}$. That is, let $\mathcal{S} \sim \mathcal{D}^n$. Define the γ-Lipschitz Loss function

$$\psi(x) := \begin{cases} 1 & x < 0 \\ 1 - \frac{x}{\gamma} & 0 \leq x \leq \gamma \\ 0 & x > \gamma \end{cases}$$

Let $\varepsilon := (\frac{8}{\gamma} + 1)\sqrt{\frac{\ln(4/\delta)}{2n}}$. Then with probability at least $1 - \delta$ over the choice of $\mathcal{S}$, for all $f \in \mathcal{F}$ we have

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}}[yf(x) \leq 0] \leq \frac{1}{n}\sum_{i=1}^{n} \psi(y_i f(\mathbf{x}_i)) + \frac{2}{\gamma}\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + \varepsilon$$

In this paper, our interest is in bounding this $\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})$ term under reasonable assumptions on $\mathcal{F}$. We specifically consider two hypothesis classes defined over a set of m kernels. First, we consider optimal kernel sum classification, where we must use the sum of all m given kernels:

$$\begin{aligned} \mathcal{F}_\Sigma := \Bigg\{ & \mathbf{x} \mapsto \sum_{i=1}^{n} \alpha_i y_i k_\Sigma(\mathbf{x}, \mathbf{x}_i), \mathbf{x}_i \in \mathcal{X}, y_i \in \{-1, 1\}, \\ & \boldsymbol{\alpha} = \text{DualSVM}(\tilde{\boldsymbol{K}}_\Sigma), \\ & \boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t \leq B^2 \ \ \forall t \in [m] \Bigg\} \end{aligned} \tag{6}$$

Second, we consider optimal kernel subset classification, where we are allowed to use the sum of any subset of the m given kernels:

$$\begin{aligned} \mathcal{F}_\mathcal{P} := \Bigg\{ & \mathbf{x} \mapsto \sum_{i=1}^{n} \alpha_i y_i k_\mathcal{P}(\mathbf{x}, \mathbf{x}_i), \mathbf{x}_i \in \mathcal{X}, y_i \in \{-1, 1\}, \\ & \mathcal{P} \subseteq [m], \boldsymbol{\alpha} = \text{DualSVM}(\tilde{\boldsymbol{K}}_\mathcal{P}), \\ & \boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t \leq B^2 \ \ \forall t \in [m] \Bigg\} \end{aligned} \tag{7}$$

Note that y_i is not present in Bartlett and Mendelson's hypothesis class in Equation 5, but it is in Equation 6 and

Equation 7. Regardless, $\mathcal{F}_\Sigma$ and $\mathcal{F}_\mathcal{P}$ do not allow for a more general set of $\boldsymbol{\alpha}$ vectors. This is because α_i is allowed to be both positive and negative in $\mathcal{F}$. However, in $\mathcal{F}_\Sigma$ and $\mathcal{F}_\mathcal{P}$, $\boldsymbol{\alpha}$ is a dual optimal vector. Dual Feasibility implies $\alpha_i \geq 0$. Thus, by explicitly mentioning y_i in the definitions of $\mathcal{F}_\Sigma$ and $\mathcal{F}_\mathcal{P}$, we are stating that α_i in $\mathcal{F}$ equals $\alpha_i y_i$ in $\mathcal{F}_\Sigma$ and $\mathcal{F}_\mathcal{P}$.

Initial Rademacher complexity bounds for learning with a single kernel assume that $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha} \leq B^2$ (Bartlett & Mendelson, 2002). Previous lines of work on multiple kernel learning then assume that $\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \leq C^2$ for some constant C (Cortes et al., 2010; 2009a; Sinha & Duchi, 2016; Srebro & Ben-David, 2006). We are interested in proving what values of C are reasonable. To achieve this, we assume that $\boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t \leq B^2$ for all base kernels and show that $\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma$ is indeed bounded.

In Section 3, we leverage our assumption that $\boldsymbol{\alpha}$ is optimal to build this bound on $\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma$. In Section 4, we demonstrate how our bound can augment existing techniques for bounding the Rademacher complexities of Equation 6 and Equation 7. That is, we bound $\hat{\mathfrak{R}}_\mathcal{S}(\mathcal{F}_\Sigma)$ and $\hat{\mathfrak{R}}_\mathcal{S}(\mathcal{F}_\mathcal{P})$.

3. SVM Bounds for Sums of Kernels

In this section, we leverage KKT conditions and SVM optimality to control the value of $\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha}$ as the number of kernels grows. To start, we consider a single kernel k:

Lemma 1. *Let $\boldsymbol{\alpha} = \text{DualSVM}(\tilde{\boldsymbol{K}})$ for some kernel matrix $\tilde{\boldsymbol{K}}$. Then $\|\boldsymbol{\alpha}\|_1 = \boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha}$.*

Proof. This proof follows from the KKT conditions provided in Section 2. We start by substituting Stationarity (Equation 2) into Complementary Slackness (Equation 4). For all $i \in [n]$,

$$
\begin{aligned}
0 &= \alpha_i (1 - y_i \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_i)) \\
0 &= \alpha_i \left(1 - \left(\sum_{j=1}^n \alpha_j y_j \boldsymbol{\phi}(\mathbf{x}_j)\right)^\intercal y_i \boldsymbol{\phi}(\mathbf{x}_i)\right) \\
0 &= \alpha_i \left(1 - \sum_{j=1}^n \alpha_j y_i y_j \boldsymbol{\phi}(\mathbf{x}_j)^\intercal \boldsymbol{\phi}(\mathbf{x}_i)\right) \\
0 &= \alpha_i - \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \boldsymbol{\phi}(\mathbf{x}_j)^\intercal \boldsymbol{\phi}(\mathbf{x}_i) \\
\alpha_i &= \sum_{j=1}^n \alpha_i \alpha_j [\tilde{\boldsymbol{K}}]_{i,j}
\end{aligned}
$$

We can then take the sum of both sides over all i:

$$
\begin{aligned}
\sum_{i=1}^n \alpha_i &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j [\tilde{\boldsymbol{K}}]_{i,j} \\
\|\boldsymbol{\alpha}\|_1 &= \boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha} \qquad (8)
\end{aligned}
$$

Note that $\sum_{i=1}^n \alpha_i = \|\boldsymbol{\alpha}\|_1$ since Dual Feasibility (Equation 3) tells us that $\alpha_i \geq 0$. □

Lemma 1 is mathematically meaningful since at the optimal point $\boldsymbol{\alpha}$, the Dual SVM takes objective value exactly equal to $\frac{1}{2}\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha}$. This connects the objective value at the optimal point to the term we want to control. With this in mind, we now move on to consider having two kernels k_1 and k_2.

Theorem 2. *Let $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be a dataset. Let k_1, k_2 be kernel functions. Define $k_{1+2}(\cdot,\cdot) := k_1(\cdot,\cdot) + k_2(\cdot,\cdot)$. Let $\tilde{\boldsymbol{K}}_1, \tilde{\boldsymbol{K}}_2, \tilde{\boldsymbol{K}}_{1+2}$ be their labeled kernel matrices and $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \boldsymbol{\alpha}_{1+2}$ be the corresponding Dual SVM solutions. Then we have*

$$
\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_{1+2} \boldsymbol{\alpha}_{1+2} \leq \frac{1}{3}(\boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1 \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2^\intercal \tilde{\boldsymbol{K}}_2 \boldsymbol{\alpha}_2)
$$

Furthermore,

$$
\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_{1+2} \boldsymbol{\alpha}_{1+2} \leq \frac{2}{3} \max\{\boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1 \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2^\intercal \tilde{\boldsymbol{K}}_2 \boldsymbol{\alpha}_2\} \qquad (9)
$$

Proof. First recall that if $\boldsymbol{\alpha} = \text{DualSVM}(\tilde{\boldsymbol{K}})$, then for all other dual feasible $\boldsymbol{\alpha}'$,

$$
\|\boldsymbol{\alpha}'\|_1 - \frac{1}{2}\boldsymbol{\alpha}'^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha}' \leq \|\boldsymbol{\alpha}\|_1 - \frac{1}{2}\boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha} \qquad (10)
$$

Also note that $\tilde{\boldsymbol{K}}_{1+2} = \tilde{\boldsymbol{K}}_1 + \tilde{\boldsymbol{K}}_2$. We start the proof by looking at the Dual SVM objective for k_{1+2}, and distributing over the labeled kernel matrices:

$$
\begin{aligned}
&\|\boldsymbol{\alpha}_{1+2}\|_1 - \frac{1}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_{1+2} \boldsymbol{\alpha}_{1+2} \\
&\quad = \|\boldsymbol{\alpha}_{1+2}\|_1 - \frac{1}{2}\boldsymbol{\alpha}_{1+2}^\intercal (\tilde{\boldsymbol{K}}_1 + \tilde{\boldsymbol{K}}_2) \boldsymbol{\alpha}_{1+2} \\
&\quad = \|\boldsymbol{\alpha}_{1+2}\|_1 - \frac{1}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_1 \boldsymbol{\alpha}_{1+2} \\
&\qquad\qquad - \frac{1}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_2 \boldsymbol{\alpha}_{1+2}
\end{aligned}
$$

We now introduce an extra $\|\boldsymbol{\alpha}_{1+2}\|_1$ term by adding zero. This allows us to form two expressions that look like Dual SVM Objectives.

$$
\begin{aligned}
&\|\boldsymbol{\alpha}_{1+2}\|_1 - \frac{1}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_{1+2} \boldsymbol{\alpha}_{1+2} \\
&\quad = \|\boldsymbol{\alpha}_{1+2}\|_1 - \frac{1}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_1 \boldsymbol{\alpha}_{1+2} \\
&\qquad\qquad - \frac{1}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_2 \boldsymbol{\alpha}_{1+2} \\
&\qquad\qquad + \|\boldsymbol{\alpha}_{1+2}\|_1 - \|\boldsymbol{\alpha}_{1+2}\|_1 \\
&\quad = \left(\|\boldsymbol{\alpha}_{1+2}\|_1 - \frac{1}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_1 \boldsymbol{\alpha}_{1+2}\right) \\
&\qquad + \left(\|\boldsymbol{\alpha}_{1+2}\|_1 - \frac{1}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_2 \boldsymbol{\alpha}_{1+2}\right) \\
&\qquad - \|\boldsymbol{\alpha}_{1+2}\|_1
\end{aligned}
$$

We then apply Inequality 10 to both of these parentheses:

$$\begin{aligned}\|\boldsymbol{\alpha}_{1+2}\|_1 - \frac{1}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_{1+2}\boldsymbol{\alpha}_{1+2} \\ \leq \left(\|\boldsymbol{\alpha}_1\|_1 - \frac{1}{2}\boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1\boldsymbol{\alpha}_1\right) \\ + \left(\|\boldsymbol{\alpha}_2\|_1 - \frac{1}{2}\boldsymbol{\alpha}_2^\intercal \tilde{\boldsymbol{K}}_2\boldsymbol{\alpha}_2\right) \\ - \|\boldsymbol{\alpha}_{1+2}\|_1\end{aligned}$$

Reorganizing the above equation, we get

$$\begin{aligned}2\|\boldsymbol{\alpha}_{1+2}\|_1 - \frac{1}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_{1+2}\boldsymbol{\alpha}_{1+2} \\ \leq \left(\|\boldsymbol{\alpha}_1\|_1 - \frac{1}{2}\boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1\boldsymbol{\alpha}_1\right) \\ + \left(\|\boldsymbol{\alpha}_2\|_1 - \frac{1}{2}\boldsymbol{\alpha}_2^\intercal \tilde{\boldsymbol{K}}_2\boldsymbol{\alpha}_2\right)\end{aligned} \tag{11}$$

Next, we use Lemma 1 to simplify all three expression that remain:

- $2\|\boldsymbol{\alpha}_{1+2}\|_1 - \frac{1}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_{1+2}\boldsymbol{\alpha}_{1+2} = \frac{3}{2}\|\boldsymbol{\alpha}_{1+2}\|_1$
- $\|\boldsymbol{\alpha}_1\|_1 - \frac{1}{2}\boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1\boldsymbol{\alpha}_1 = \frac{1}{2}\|\boldsymbol{\alpha}_1\|_1$
- $\|\boldsymbol{\alpha}_2\|_1 - \frac{1}{2}\boldsymbol{\alpha}_2^\intercal \tilde{\boldsymbol{K}}_2\boldsymbol{\alpha}_2 = \frac{1}{2}\|\boldsymbol{\alpha}_2\|_1$

Returning to our bound from Inequality 11, we have

$$\frac{3}{2}\|\boldsymbol{\alpha}_{1+2}\|_1 \leq \frac{1}{2}\|\boldsymbol{\alpha}_1\|_1 + \frac{1}{2}\|\boldsymbol{\alpha}_2\|_1 \tag{12}$$

Once we rearrange the constants in Inequality 12, we complete the proof. □

The constant of $\frac{2}{3}$ in Inequality 9 is advantageous. Since this ratio is below 1, we can recursively apply this theorem to get a vanishing fraction. As m increases, we should now expect $\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma$ to decrease. We formalize this notion in the following theorem, where we consider using the sum of m kernels.

Theorem 3. *Let $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be a dataset. Let $k_1, k_2, \ldots, k_m$ be kernel functions. Define $k_\Sigma(\cdot,\cdot) := \sum_{t=1}^m k_t(\cdot,\cdot)$. Let $\tilde{\boldsymbol{K}}_1, \ldots, \tilde{\boldsymbol{K}}_m, \tilde{\boldsymbol{K}}_\Sigma$ be their labeled kernel matrices and $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m, \boldsymbol{\alpha}_\Sigma$ be the corresponding Dual SVM solutions. Then we have*

$$\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \leq 3m^{-\log_2(3)} \sum_{t=1}^m \boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t$$

Furthermore,

$$\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \leq 3m^{-\log_2(3/2)} \max_{t\in[m]} \boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t$$

In the special case that m is a power of 2, we have

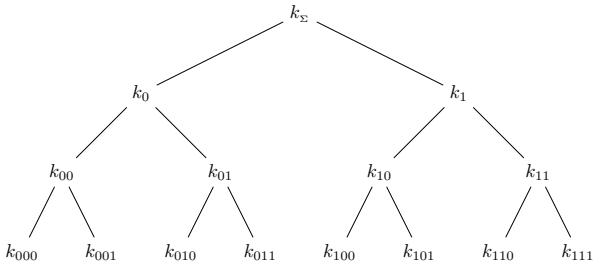
$$\begin{aligned}\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma &\leq m^{-\log_2(3)} \sum_{t=1}^m \boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t \\ &\leq m^{-\log_2(3/2)} \max_{t\in[m]} \boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t\end{aligned}$$

Proof sketch. We provide an intuitive proof for $m = 8$. The full proof is in Appendix B.

Since m is a power of two, we can label each of the base kernels with length $\ell = \log_2(m) = 3$ bitstrings:

k_{000} k_{001} k_{010} k_{011} k_{100} k_{101} k_{110} k_{111}

Then, for each pair of kernels that differ only in the last digit, define a new kernel as their sum. For instance, define $k_{10}(\cdot,\cdot) := k_{100}(\cdot,\cdot) + k_{101}(\cdot,\cdot)$. Repeat this process all the way to the root node.



By Theorem 3, we know that

$$\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \leq \frac{1}{3}(\boldsymbol{\alpha}_0^\intercal \tilde{\boldsymbol{K}}_0 \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1 \boldsymbol{\alpha}_1)$$

Going down one level, by applying Theorem 3 again, we know that

$$\boldsymbol{\alpha}_0^\intercal \tilde{\boldsymbol{K}}_0 \boldsymbol{\alpha}_0 \leq \frac{1}{3}(\boldsymbol{\alpha}_{00}^\intercal \tilde{\boldsymbol{K}}_{00} \boldsymbol{\alpha}_{00} + \boldsymbol{\alpha}_{01}^\intercal \tilde{\boldsymbol{K}}_{01} \boldsymbol{\alpha}_{01})$$

Therefore, by similarly applying Theorem 3 to $\boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1 \boldsymbol{\alpha}_1$, we can combine these claims:

$$\begin{aligned}\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \leq \left(\frac{1}{3}\right)^2 (\boldsymbol{\alpha}_{00}^\intercal \tilde{\boldsymbol{K}}_{00} \boldsymbol{\alpha}_{00} + \boldsymbol{\alpha}_{01}^\intercal \tilde{\boldsymbol{K}}_{01} \boldsymbol{\alpha}_{01} + \\ \boldsymbol{\alpha}_{10}^\intercal \tilde{\boldsymbol{K}}_{10} \boldsymbol{\alpha}_{10} + \boldsymbol{\alpha}_{11}^\intercal \tilde{\boldsymbol{K}}_{11} \boldsymbol{\alpha}_{11})\end{aligned}$$

We can then continue until all 8 kernels are included:

$$\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \leq \left(\frac{1}{3}\right)^3 \sum_{t=1}^m \boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t$$

Note that the exponent of $\frac{1}{3}$ is the depth of the tree, equivalent to the length ℓ of our bitstring labels. In the general case, we have

$$\begin{aligned}\boldsymbol{\alpha}_{\Sigma}^{\intercal}\tilde{\boldsymbol{K}}_{\Sigma}\boldsymbol{\alpha}_{\Sigma} &\leq \left(\frac{1}{3}\right)^{\log_2(m)}\sum_{t=1}^{m}\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t \\ &= m^{-\log_2(3)}\sum_{t=1}^{m}\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t\end{aligned}$$

This completes the analysis if m is a power of 2. If we do not have an exact power of two number of kernels, then our tree has depth $\ell - 1$ for some leaves. Therefore, we place a floor function around $\log_2(3)$:

$$\begin{aligned}\boldsymbol{\alpha}_{\Sigma}^{\intercal}\tilde{\boldsymbol{K}}_{\Sigma}\boldsymbol{\alpha}_{\Sigma} &\leq \left(\frac{1}{3}\right)^{\lfloor\log_2(m)\rfloor}\sum_{t=1}^{m}\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t \\ &\leq 3\left(\frac{1}{3}\right)^{\log_2(m)}\sum_{t=1}^{m}\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t \\ &= 3m^{-\log_2(3)}\sum_{t=1}^{m}\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t\end{aligned}$$

To achieve the final result, we bound the summation with

$$\sum_{t=1}^{m}\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t \leq m\max_{t\in[m]}\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t$$

and simplify the resulting expression. □

We take a moment to reflect on this result. It has been well established that the generalization error of kernel classifiers depends on $\boldsymbol{\alpha}_{\Sigma}^{\intercal}\tilde{\boldsymbol{K}}_{\Sigma}\boldsymbol{\alpha}_{\Sigma}$ (Cortes et al., 2009c; 2012; 2013; Srebro & Ben-David, 2006; Sinha & Duchi, 2016). Theorem 3 shows that this term actually decreases in the number of kernels. In the next section, we show how this theorem translates into generalization error results.

4. Rademacher Bounds

In this section we apply Theorem 3 to bound the Rademacher complexity of learning with sums of kernels. To better parse and understand these bounds, we make two common assumptions:

- Each base kernel has a bounded Dual SVM solution:

$$\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t \leq B^2 \quad \forall t \in [m]$$

- Each vector has a bounded ℓ_2 norm in each feature space:

$$k_t(\mathbf{x}_i, \mathbf{x}_i) \leq R^2 \quad \forall t \in [m], i \in [n]$$

The classical bound in (Bartlett & Mendelson, 2002) on the Rademacher complexity of kernel functions looks at a *single* kernel, and provides the bound

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) \leq \frac{BR}{\sqrt{n}}$$

where the hypothesis class $\mathcal{F}$ is defined in Equation 5. Our bounds are on the order of $\frac{BR}{\sqrt{n}}m^{0.208}$. That is, when moving from one kernel to many kernels, we pay sublinearly in the number of kernels.

We first see this with our bound on the Rademacher complexity of the kernel sum hypothesis class $\hat{\mathcal{R}}_{\mathcal{S}}(\mathcal{F}_{\Sigma})$ defined in Equation 6:

Theorem 4. *Let $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be a dataset. Let $k_1, \ldots, k_m$ be kernel functions. Define $k_{\Sigma}(\cdot,\cdot) := \sum_{t=1}^{m} k_t(\cdot,\cdot)$. Let $\tilde{\boldsymbol{K}}_1, \ldots, \tilde{\boldsymbol{K}}_m, \tilde{\boldsymbol{K}}_{\Sigma}$ be their labeled kernel matrices and $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m, \boldsymbol{\alpha}_{\Sigma}$ be the corresponding Dual SVM solutions. Then,*

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}_{\Sigma}) \leq \frac{1}{n}\sqrt{3m^{-\log_2(3)}\left(\sum_{t=1}^{m}\operatorname{Tr}[\tilde{\boldsymbol{K}}_t]\right)\sum_{t=1}^{m}\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t}$$

Furthermore, if we assume that $\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t \leq B^2$ and $k_t(\mathbf{x}_i, \mathbf{x}_i) \leq R^2$ for all $t \in [m]$ and $i \in [n]$, then we have

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}_{\Sigma}) \leq \frac{BR}{\sqrt{n}}\sqrt{3m^{(1-\log_2(3/2))}} \in O\left(\frac{BRm^{0.208}}{\sqrt{n}}\right)$$

Our proof parallels that of Lemma 22 in (Bartlett & Mendelson, 2002), and a full proof is in Appendix C. The key difference between Bartlett and Mendelson's proof and ours is the assumption that $\boldsymbol{\alpha}$ is optimal, allowing us to apply Theorem 3.

Next, we consider learning which kernels to sum. In this setting, we allow an algorithm to pick any subset of kernels to sum, but require that Kernel SVM is used for prediction. This is described by the hypothesis class $\mathcal{F}_{\mathcal{P}}$ defined in Equation 7. Because the algorithm can pick any arbitrary subset, we are intuitively bounded by the worst risk over all subsets of kernels. Specifically, Theorem 4 suggests that the risk of $\mathcal{F}_{\mathcal{P}}$ is bounded by the risk of a subset with size m. That is, the risk of $\mathcal{F}_{\mathcal{P}}$ is bounded by the risk of using all kernels. Our next theorem makes this intuition precise, because we only pay an asymptotic factor of $\sqrt{\ln(m)}$ more when considering all possible subsets of kernels instead of only one subset of kernels.

Theorem 5. *Let $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be a dataset. Let $k_1, \ldots, k_m$ be kernel functions. Consider any $\mathcal{P} \subseteq [m]$. Define $k_{\mathcal{P}}(\cdot,\cdot) := \sum_{t\in\mathcal{P}} k_t(\cdot,\cdot)$. Let $\tilde{\boldsymbol{K}}_1, \ldots, \tilde{\boldsymbol{K}}_m, \tilde{\boldsymbol{K}}_{\mathcal{P}}$ be their labeled kernel matrices and $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m, \boldsymbol{\alpha}_{\mathcal{P}}$ be the*

corresponding Dual SVM solutions. Assume $k_t(\mathbf{x}_i, \mathbf{x}_i) \leq R^2$ *and* $\boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t \leq B^2$ *for all* $t \in [m]$ *and* $i \in [n]$*. Then,*

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}_{\mathcal{P}}) \leq \frac{BR\sqrt{3e\eta_0 \, m^{(1-\log_2(3/2))} \lceil \ln(m) \rceil}}{\sqrt{n}} \in O\left(\frac{BRm^{0.208}\sqrt{\ln(m)}}{\sqrt{n}}\right)$$

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

If we tried to build this bound with the classical analytical method found in Lemma 22 of (Bartlett & Mendelson, 2002), we would have to deal with a difficult supremum over the 2^m distinct choices of kernels. This would inflate the bound by a multiplicative factor of $\sqrt{2^m} = 2^{m-1}$. However, our proof instead follows that of Theorem 1 in (Cortes et al., 2009c). This more complicated proof method allows us to pay a factor of $\sqrt{ln(m)}$ to separate supremum over the choice of kernels and the expectation over the $\boldsymbol{\sigma}$ vector. This separation allows us to invoke Theorem 3. However, this proof technique also prevents us from building a claim as general as Theorem 4, instead only providing bounds using B and R. The full proof is found in Appendix D.

5. Bounds for Non-separable Data

Recall the Primal and Dual SVMs with ℓ_2 slack variables from Section 2. Now we show that if $C = \frac{1}{2}$, then all the other bounds hold using non-separable SVM instead of the separable one. We achieve this by mirroring Theorem 2, which is used by all other results in this paper.

Theorem 6. *Let* $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ *be a dataset. Let* k_1, k_2 *be kernel functions. Define* $k_{1+2}(\cdot,\cdot) := k_1(\cdot,\cdot) + k_2(\cdot,\cdot)$*. Let* $\tilde{\boldsymbol{K}}_1, \tilde{\boldsymbol{K}}_2, \tilde{\boldsymbol{K}}_{1+2}$ *be their labeled kernel matrices and* $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_{1+2}$ *be the corresponding Dual SVM solutions with parameter* $C = \frac{1}{2}$*. Then we have*

$$\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_{1+2} \boldsymbol{\alpha}_{1+2} \leq \frac{1}{3}(\boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1 \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2^\intercal \tilde{\boldsymbol{K}}_2 \boldsymbol{\alpha}_2)$$

Furthermore,

$$\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_{1+2} \boldsymbol{\alpha}_{1+2} \leq \frac{2}{3} \max\{\boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1 \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2^\intercal \tilde{\boldsymbol{K}}_2 \boldsymbol{\alpha}_2\}$$

The proof mirrors the proof of Theorem 3, except for some careful book keeping for the slack vectors $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$, and $\boldsymbol{\xi}_{1+2}$. Again, it is the KKT conditions that allow us to bound and compare the $\boldsymbol{\xi}$ vectors of the three Dual SVM problems. The full proof is in Appendix A.

With Theorem 6, we can reproduce all other results without any changes to the original proofs. Further, this sort of bound on C being a constant is common in learning theory literature such as PAC Bayes (McAllester, 2007).

6. Experiment

We show some experimental results that verify our core theorem, i.e. Theorem 3. Our experiment uses 8 fixed kernels from several kernels families. We have 5 radial basis kernels, 1 linear kernel, 1 polynomial kernel, and 1 cosine kernel. All our data is generated from a mixture of 4 Gaussians. Two of the Gaussians generate the positive class while the other 2 generate the negative class.

We generate $n = 300$ samples in $\mathbb{R}^{50}$. For each of the 8 base kernels, we solve the Dual Kernel SVM problem, and empirically verify that $\boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t \leq 320 = B^2$.

Then, we arbitrarily permute the kernel matrices. We solve the Dual Kernel SVM problem with the first kernel matrix denoted as $\tilde{\boldsymbol{K}}_{\Sigma,1}$. Then we solve the SVM with the sum of the first two kernels, denoted as $\tilde{\boldsymbol{K}}_{\Sigma,2}$, and so on until we sum all 8 kernels. Let $\boldsymbol{\alpha}_{\Sigma,m}$ denote the dual solution vector corresponding sum of the first m of the 8 kernels. That is,

$$\boldsymbol{\alpha}_{\Sigma,m} := \text{DualSVM}(\tilde{\boldsymbol{K}}_{\Sigma,m}) = \text{DualSVM}\left(\sum_{t=1}^{m} \tilde{\boldsymbol{K}}_t\right)$$

After solving each SVM problem, we keep track of the value of $\boldsymbol{\alpha}_{\Sigma,m}^\intercal \tilde{\boldsymbol{K}}_{\Sigma,m} \boldsymbol{\alpha}_{\Sigma,m}$ value. We then plot this value against the two bounds provided by Theorem 3:

$$\begin{aligned}\boldsymbol{\alpha}_{\Sigma,m}^\intercal \tilde{\boldsymbol{K}}_{\Sigma,m} \boldsymbol{\alpha}_{\Sigma,m} &\leq m^{-\log_2(3)} \sum_{t=1}^{m} \boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t \\ &\leq m^{-\log_2(3/2)} B^2\end{aligned}$$

Figure 1 shows the difference between the true $\boldsymbol{\alpha}_{\Sigma,m}^\intercal \tilde{\boldsymbol{K}}_{\Sigma,m} \boldsymbol{\alpha}_{\Sigma,m}$ and the two bounds above. We can observe that the true curve decreases roughly at the same rate as our bounds.

7. Conclusion

Here we discuss possible directions to extend our work. First, in the context of classification with kernel sums, we are not aware of any efficient and theoretically sound algorithms for learning which kernels to sum. Additionally, we believe that optimality conditions such as KKT are necessary to build meaningful lower bounds in this setting.

One could also analyze the sample complexity of kernel products. This idea is experimentally considered by (Duvenaud et al., 2013). This problem is notably more difficult since it requires understanding the Hadamard product of kernel matrices.

More generally, there is little existing work that leverages optimality conditions to justify assumptions made in learning problems. In this paper, KKT tells us that we can control

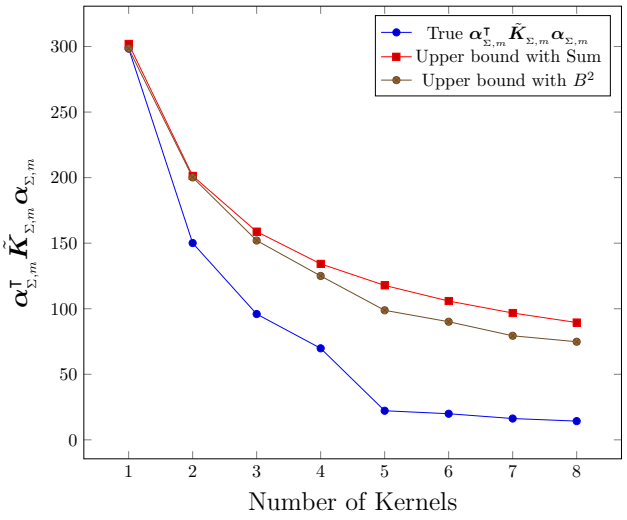


Figure 1. Empirical value and bounds of $\boldsymbol{\alpha}_{\Sigma,m}^\intercal \tilde{\boldsymbol{K}}_{\Sigma,m} \boldsymbol{\alpha}_{\Sigma,m}$ in our experiment. The blue curve is the empirical $\boldsymbol{\alpha}_{\Sigma,m}^\intercal \tilde{\boldsymbol{K}}_{\Sigma,m} \boldsymbol{\alpha}_{\Sigma,m}$. The brown curve corresponds to $m^{-\log_2(3)} \sum_{t=1}^m \boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t$. The red curve corresponds to $m^{-\log_2(3/2)} B^2$.

the quantity $\boldsymbol{\alpha}_\Sigma \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma$, justifying the assumptions made in prior work (Cortes et al., 2009c; Srebro & Ben-David, 2006; Sinha & Duchi, 2016). We believe that this overall idea is general and applies to other convex optimization problems and classes of representer theorem problems, as well as other learning problems.

References

Argyriou, A., Micchelli, C. A., and Pontil, M. Learning convex combinations of continuously parameterized basic kernels. In *International Conference on Computational Learning Theory*, pp. 338–352. Springer, 2005.

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

Balcan, M.-F., Blum, A., and Vempala, S. Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning*, 65(1):79–94, 2006.

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

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Chen, Y., Gupta, M. R., and Recht, B. Learning kernels from indefinite similarities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 145–152. ACM, 2009.

Cortes, C., Mohri, M., and Rostamizadeh, A. L 2 regularization for learning kernels. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 109–116. AUAI Press, 2009a.

Cortes, C., Mohri, M., and Rostamizadeh, A. Learning nonlinear combinations of kernels. In *Advances in neural information processing systems*, pp. 396–404, 2009b.

Cortes, C., Mohri, M., and Rostamizadeh, A. New generalization bounds for learning kernels. *arXiv preprint arXiv:0912.3309*, 2009c.

Cortes, C., Mohri, M., and Rostamizadeh, A. Generalization bounds for learning kernels. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 247–254. Omnipress, 2010.

Cortes, C., Mohri, M., and Rostamizadeh, A. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(Mar):795–828, 2012.

Cortes, C., Kloft, M., and Mohri, M. Learning kernels using local Rademacher complexity. In *Advances in neural information processing systems*, pp. 2760–2768, 2013.

Daumé III, H. A course in machine learning. *Publisher, ciml. info*, pp. 5–73, 2012.

Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B., and Ghahramani, Z. Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302.4922*, 2013.

Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

Gönen, M. and Alpaydın, E. Multiple kernel learning algorithms. *Journal of machine learning research*, 12(Jul): 2211–2268, 2011.

Kakade, S. M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pp. 793–800, 2009.

Kivinen, J., Smola, A. J., and Williamson, R. C. Online learning with kernels. *IEEE transactions on signal processing*, 52(8):2165–2176, 2004.

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

McAllester, D. Generalization bounds and consistency. In BakIr, G., Hofmann, T., Schölkopf, B., Smola, A. J., Taskar, B., and Vishwanathan, S. (eds.), *Predicting structured data*, pp. 247–261. MIT press, 2007.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shawe-Taylor, J., Cristianini, N., et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Sinha, A. and Duchi, J. C. Learning kernels with random features. In *Advances in Neural Information Processing Systems*, pp. 1298–1306, 2016.

Soentpiet, R. et al. *Advances in kernel methods: support vector learning*. MIT press, 1999.

Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkopf, B. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7(Jul):1531–1565, 2006.

Srebro, N. and Ben-David, S. Learning bounds for support vector machines with learned kernels. In *International Conference on Computational Learning Theory*, pp. 169–183. Springer, 2006.