# Optimality Implies Kernel Sum Classifiers are Statistically Efficient


Raphael Arkady Meyer  Jean Honorio

Computer Science Department, Purdue University


## Introduction and Optimization

### Background

**Statistical Learning Theory + Optimization**

- Generalization proofs typically state that all feasible estimators generalize well
- This includes low-accuracy estimators we do not care about
- Proofs often make stringent assumptions on the data distribution
- We combine Optimization and Statistical Learning Theory to prove that *optimal estimators* generalize well
- We justify common assumptions made in the Multiple Kernel Learning literature

**Multiple Kernel Learning**

- Given $m$ kernels $k_1, \dots, k_m$ and dataset $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$
- An estimator picks $\theta_1, \dots, \theta_m$ and $\boldsymbol{\alpha}$
- Define combined kernel $k_\Sigma(\cdot,\cdot) = \sum_{t=1}^m \theta_t k_t(\cdot,\cdot)$
- Predict with $y(\mathbf{x}|\tilde{\boldsymbol{K}}_\Sigma, \boldsymbol{\alpha}) = \sum_{i=1}^n \alpha_i k_\Sigma(\mathbf{x}, \mathbf{x}_i)$

**Our Approach**

- Binary Classification: $y_i \in \{-1, +1\}$
- $\boldsymbol{\alpha}$ is optimal in a Support Vector Machine
- Control generalization error of $k_\Sigma$ with the error of $k_1, \dots, k_m$

### Optimization-Based Results

**Lemma of One Kernel**

Let $\boldsymbol{\alpha}$ be the dual-optimal vector for labeled kernel matrix $\tilde{\boldsymbol{K}}$. Then, by combining the *Stationarity*, *Complementary Slackness*, and *Dual Feasibility* KKT conditions, we find that

$$\|\boldsymbol{\alpha}\|_1 = \boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha}$$

**Theorem of Two Kernels: Adding Kernels Reduces Complexity**

Let $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ be the dual-optimal vectors for labeled kernel matrices $\tilde{\boldsymbol{K}}_1$ and $\tilde{\boldsymbol{K}}_2$. Let $\boldsymbol{\alpha}_{1+2}$ be the dual-optimal vector for labeled kernel matrix $\tilde{\boldsymbol{K}}_{1+2} := \tilde{\boldsymbol{K}}_1 + \tilde{\boldsymbol{K}}_2$. Then, following from *the prior lemma*, *the optimality of* $\boldsymbol{\alpha}_{1+2}$, and some algebra, we have

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

**Theorem of Many Kernels: Adding Many Kernels Greatly Reduces Complexity**

Let $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_m$ be the dual-optimal vectors for labeled kernel matrices $\tilde{\boldsymbol{K}}_1, \dots, \tilde{\boldsymbol{K}}_m$. Let $\boldsymbol{\alpha}_\Sigma$ be the dual optimal vector for labeled kernel matrix $\tilde{\boldsymbol{K}}_\Sigma := \sum_{t=1}^m \tilde{\boldsymbol{K}}_t$. Then, *by repeatedly applying the prior lemma*, we find

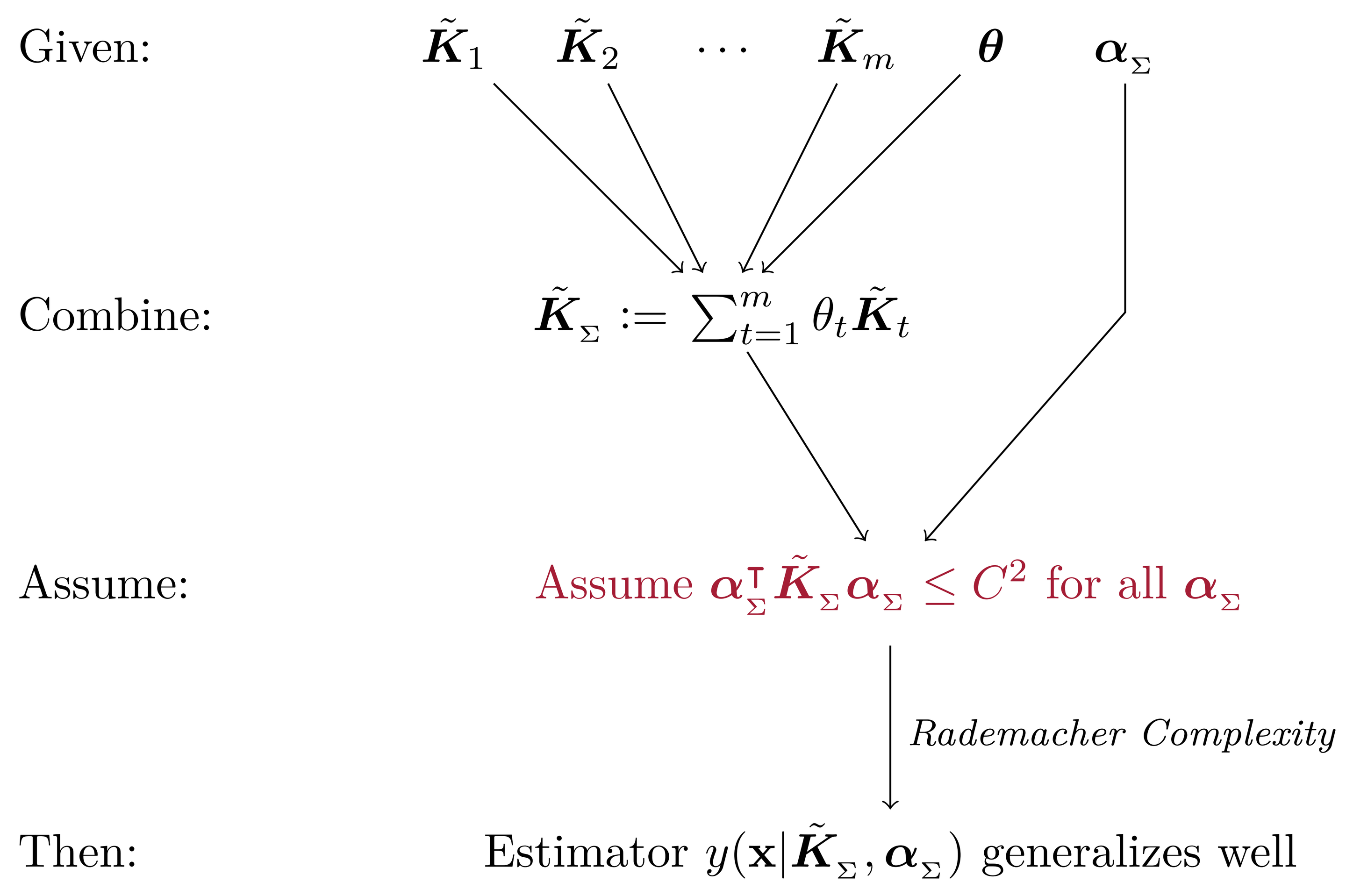
$$\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \le 3m^{-\log_2(3)} \sum_{t=1}^m \boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t$$

Furthermore, if we assume that $\boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t \le B^2$, then

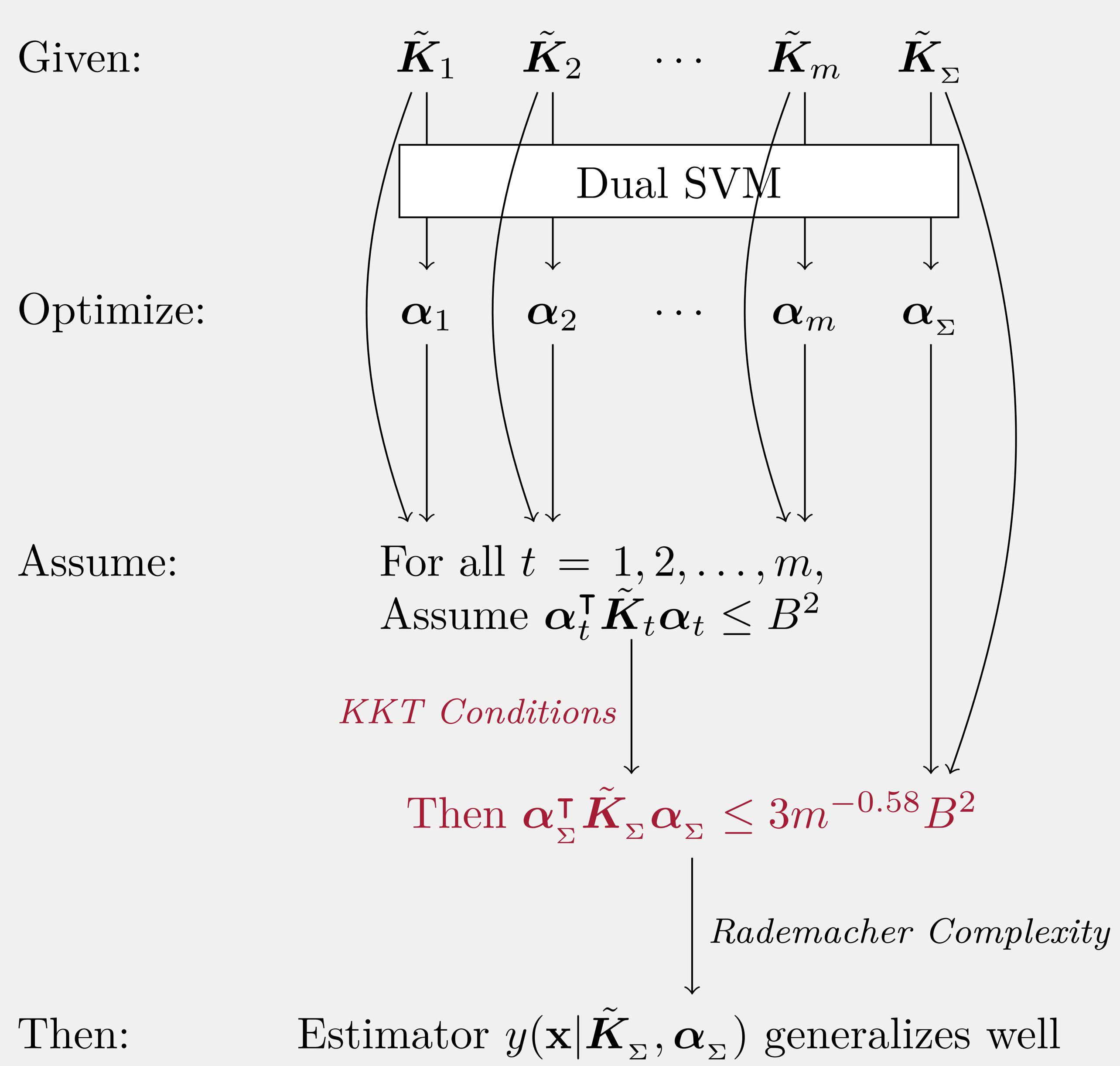
$$\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \le 3m^{-\log_2(3/2)} B^2$$

## Main Theorem and Context

### Template of Prior Works




### Our Optimality Assumption




## Conclusions

### Learning Theory Results

**Support Vector Machines Styles**

- We consider the standard SVM and a L2-penalized SVM for nonseparable data:

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

(a) Primal SVM Problem

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

(b) Dual SVM Problem

Figure 1. Primal and Dual SVM Problems. The L2 penalties are in gray.

- We prove statistical efficiency for standard SVM and $C = \frac{1}{2}$ in the L2-SVM

**Ways to Combine Kernels Together**

- Our core theorem complements existing Rademacher Complexity proofs
- Generalization error is bounded by the Rademacher Complexity $\hat{\mathcal{R}}(\mathcal{F})$:

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

- Different proofs consider different ways to combine kernels:

1. Kernel Sums: If all $\theta_t = 1$, then
$$\hat{\mathcal{R}}(\mathcal{F}) = O\left(\frac{BRm^{0.208}}{\sqrt{n}}\right)$$
2. Kernel Subsets: If all $\theta_t \in \{0, 1\}$, then
$$\hat{\mathcal{R}}(\mathcal{F}) = O\left(\frac{BRm^{0.208} \cdot \sqrt{\ln(m)}}{\sqrt{n}}\right)$$
3. Convex Combinations*: If we have $\theta_t \in \{0\} \cup \left[\frac{10}{m}, 1\right]$ and $\sum_{t=1}^m \theta_t = 1$, then
$$\hat{\mathcal{R}}(\mathcal{F}) = O\left(\frac{BRm\sqrt{\ln(m)}}{\sqrt{n}}\right)$$

### Table of Constants

| Variable | Meaning |
|---|---|
| $n$ | Number of Samples |
| $i, j$ | Index of a Sample |
| $m$ | Number of Kernels |
| $t$ | Index of a Kernel |
| $\tilde{\boldsymbol{K}}$ | Labeled Kernel Matrix (i.e. $[\tilde{\boldsymbol{K}}]_{i,j} := y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$) |
| $\boldsymbol{\alpha}_t$ | Dual Solution Vector for SVM with $\tilde{\boldsymbol{K}}_t$ |
| $B^2$ | Upper Bound for all $\boldsymbol{\alpha}_t^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_t$ |
| $R^2$ | Upper Bound for all $k_t(\mathbf{x}_i, \mathbf{x}_i) = \|\boldsymbol{\phi}(\mathbf{x}_i)\|_2^2$ |