Introduction and Optimization

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 as kernels $k_1, \ldots, k_m$, and dataset $(x_1, y_1), \ldots, (x_n, y_n)$.
- An estimator picks $\alpha_1, \ldots, \alpha_m$ and $\alpha$.
- Define combined kernel $k_t(x) = \sum_{i=1}^m \theta_t k_i(x)$.
- Predict with $y(x; K, \alpha) = \sum_{i=1}^m \alpha_i k_i(x, x_i)$.

Our Approach

- Binary Classification: $y_t \in \{-1, +1\}$.
- $\alpha$ is optimal in a Support Vector Machine.
- Control generalization error of $k_t$ with the error of $k_1, \ldots, k_m$.

Optimization-Based Results

Lemma of One Kernel

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

$$\alpha = -\alpha^T K \alpha$$

Theorem of Two Kernels: Adding Kernels Reduces Complexity

Let $\alpha_1$ and $\alpha_2$ be the dual-optimal vectors for labeled kernel matrices $K_1$ and $K_2$. Let $\alpha_{1,2}$ be the dual-optimal vector for labeled kernel matrix $K_{1,2} = K_1 + K_2$. Then, following from the prior lemma, the optimality of $\alpha_{1,2}$ and some algebra, we have

$$\alpha_{1,2}^T K_{1,2} \alpha_{1,2} \leq \frac{1}{4} \alpha_1^T K_1 \alpha_1 + \alpha_2^T K_2 \alpha_2$$

Theorem of Many Kernels: Adding Many Kernels Greatly Reduces Complexity

Let $\alpha_1, \ldots, \alpha_m$ be the dual-optimal vectors for labeled kernel matrices $K_1, \ldots, K_m$. Let $\alpha_e$ be the dual optimal vector for labeled kernel matrix $K_e = \sum_{i=1}^m K_i$. Then, by repeatedly applying the prior lemma, we find

$$\alpha_e^T K_e \alpha_e \leq m^{-0.5} B^2$$

Furthermore, if we assume that $\alpha_e^T K_e \alpha_e \leq B^2$, then

$$\alpha_e^T K_e \alpha_e \leq m^{-1} \ln(2)$$

Main Template and Context

### Template of Prior Works

**Given:**

$K_1, K_2, \ldots, K_m, \alpha_e$

**Optimize:**

$\alpha_1, \alpha_2, \ldots, \alpha_m, \alpha_e$

**Assume:**

For all $t = 1, 2, \ldots, m$, Assume $\alpha_e^T K_e \alpha_e \leq B^2$

**KKT Conditions**

Then $\alpha_e^T K_e \alpha_e \leq 3m^{-0.5} B^2$

**Rademacher Complexity**

Then estimator $y(x; K_e, \alpha_e)$ generalizes well.

### Learning Theory Results

**Support Vector Machines Styles**

- We consider the standard SVM and a $L_2$-penalized SVM for nonseparable data:

$$\min \frac{1}{n} \sum_{i=1}^n \max \{0, 1 - y_i w^T x_i\}$$

**Ways to Combine Kernels Together**

- Our core theorem complements existing Rademacher Complexity proofs.
- Generalization error is bounded by the Rademacher Complexity $R(F)$:

$$R(F) = \hat{R} = \hat{R}(F) = \frac{\hat{R}(F)}{\sqrt{m}}$$

**Different proofs consider different ways to combine kernels:**

1. **Kernel Sums:** If all $\theta_t = 1$,

$$R(F) = O\left(\frac{B B_0 \ln m}{\sqrt{m}}\right)$$

2. **Kernel Subsets:** If all $\theta_t \in \{0, 1\}$,

$$R(F) = O\left(\frac{B B_0 \ln m}{\sqrt{m}}\right)$$

3. **Convex Combinations:** If we have $\theta_t \in \{0, 1\}$ and $\sum_{i=1}^m \theta_i = 1$, then

$$R(F) = O\left(\frac{B B_0 \ln m}{\sqrt{m}}\right)$$

### Table of Constants

<table>
<thead>
<tr>
<th>Variable</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>Number of Samples</td>
</tr>
<tr>
<td>$i$</td>
<td>Index of a Sample</td>
</tr>
<tr>
<td>$m$</td>
<td>Number of Kernels</td>
</tr>
<tr>
<td>$K$</td>
<td>Kernel Matrix</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Dual Solution Vector for SVM with $K$</td>
</tr>
<tr>
<td>$R^*$</td>
<td>Upper bound for all $K_{ij}$</td>
</tr>
<tr>
<td>$R^a$</td>
<td>Upper bound for all $k_i(x, x_j) =</td>
</tr>
</tbody>
</table>