A. Non-Separable Proof of Two Kernels (Theorem 6)

In this section, we prove a theorem that mirrors that of Theorem 2, but with the ℓ_2 slack SVM. First, we state the KKT conditions for the slack SVM. Let $\mathbf{r}$ be the dual variables associated with the primal $\boldsymbol{\xi} \succeq 0$ constraints. Then, we have 8 conditions:

1. $1 - \xi_i - y_i \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_i) \leq 0 \ \ \forall i \in [n]$ *(Primal Feasibility 1)*
2. $\xi_i \geq 0 \ \ \forall i \in [n]$ *(Primal Feasibility 2)*
3. $\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \boldsymbol{\phi}(\mathbf{x}_i)$ *(Stationarity 1)*
4. $\mathbf{r} = C\boldsymbol{\xi} - \boldsymbol{\alpha}$ *(Stationarity 2)*
5. $\alpha_i \geq 0 \ \ \forall i \in [n]$ *(Dual Feasibility 1)*
6. $r_i \geq 0 \ \ \forall i \in [n]$ *(Dual Feasibility 2)*
7. $\alpha_i(1 - \xi_i - y_i \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_i)) = 0 \ \ \forall i \in [n]$ *(Complementary Slackness 1)*
8. $r_i \xi_i = 0 \ \ \forall i \in [n]$ *(Complementary Slackness 2)*

We also provide two preliminary lemmas before proving the main theorem.

Lemma 2. *Let $\boldsymbol{\alpha}, \boldsymbol{\xi}$ be the optimal solution to the ℓ_2 Slack Dual SVM problem with parameter C. Then, $\boldsymbol{\xi} = \frac{1}{C}\boldsymbol{\alpha}$. This also implies $\boldsymbol{\alpha}^\intercal \boldsymbol{\xi} = C\|\boldsymbol{\xi}\|_2^2$.*

Proof. First we substitute Stationarity 2 into Complementary Slackness 2:

$$\begin{aligned} r_i \xi_i &= 0 \\ (C\xi_i - \alpha_i)\xi_i &= 0 \\ C\xi_i^2 &= \alpha_i \xi_i \end{aligned}$$

That is, when $\xi_i \neq 0$, we know that $\xi_i = \frac{\alpha_i}{C}$. This allows us to conclude that $\xi_i \leq \frac{\alpha_i}{C}$, since both α_i and C are nonnegative. The dual problem has constraint $\alpha_i \leq C\xi_i$, which is equivalent to $\xi_i \geq \frac{\alpha_i}{C}$. Hence ξ_i is both upper and lower bounded by $\frac{\alpha_i}{C}$. Therefore, $\xi_i = \frac{\alpha_i}{C}$. $\square$

Lemma 3. *Let $\boldsymbol{\alpha}, \boldsymbol{\xi}$ be the optimal solution to the ℓ_2 Slack Dual SVM problem on input $\tilde{\boldsymbol{K}}$ with parameter C. Then $\|\boldsymbol{\alpha}\|_1 = \boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha} + C\|\boldsymbol{\xi}\|_2^2$.*

Proof. First substitute Stationarity 1 into Complementary Slackness 1:

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

Then, we sum up over all $i \in [n]$:

$$\begin{aligned}
\sum_{i=1}^{n} \alpha_i &= \sum_{i=1}^{n} \alpha_i \xi_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j [\tilde{\boldsymbol{K}}]_{i,j} \\
\|\boldsymbol{\alpha}\|_1 &= \boldsymbol{\alpha}^\intercal \boldsymbol{\xi} + \boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha} \\
\|\boldsymbol{\alpha}\|_1 &= C\|\boldsymbol{\xi}\|_2^2 + \boldsymbol{\alpha}^\intercal \tilde{\boldsymbol{K}} \boldsymbol{\alpha}
\end{aligned}$$

□

Now we prove the main theorem:

Theorem 6 Restated. *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,

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

Proof. We start with the dual objective for k_{1+2}:

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

By applying Lemma 3 and some algebra, we have three useful equations:

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

Applying these equations, we continue our inequality from before,

$$\frac{3}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_{1+2}\boldsymbol{\alpha}_{1+2} + (2C-1)\|\boldsymbol{\xi}_{1+2}\|_2^2 \leq \left(\frac{1}{2}\boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1\boldsymbol{\alpha}_1 + \frac{2C-1}{2}\|\boldsymbol{\xi}_1\|_2^2\right) + \left(\frac{1}{2}\boldsymbol{\alpha}_2^\intercal \tilde{\boldsymbol{K}}_2\boldsymbol{\alpha}_2 + \frac{2C-1}{2}\|\boldsymbol{\xi}_2\|_2^2\right)$$

$$\frac{3}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_{1+2}\boldsymbol{\alpha}_{1+2} \leq \frac{1}{2}\left(\boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2^\intercal \tilde{\boldsymbol{K}}_2\boldsymbol{\alpha}_2\right) + \frac{2C-1}{2}\left(\|\boldsymbol{\xi}_1\|_2^2 + \|\boldsymbol{\xi}_2\|_2^2 - 2\|\boldsymbol{\xi}_{1+2}\|_2^2\right)$$

$$\frac{3}{2}\boldsymbol{\alpha}_{1+2}^\intercal \tilde{\boldsymbol{K}}_{1+2}\boldsymbol{\alpha}_{1+2} = \frac{1}{2}\left(\boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2^\intercal \tilde{\boldsymbol{K}}_2\boldsymbol{\alpha}_2\right) + 0$$

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

In the second to last line, we recall that $C = \frac{1}{2}$, which implies $2C - 1 = 0$. □

B. Proof of Many Kernels (Theorem 3)

Theorem 3 Restated. *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

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

Proof. Let $\ell := \lceil \log_2(m) \rceil$ be the length of labels we give our base kernels. Now, rename each kernel k_t with the length ℓ bitstring representation of the number t. For instance, if $\ell = 4$ then we rename k_6 to k_{0110}. For every length $\ell - 1$ bitstring $b_0 b_1 \ldots b_{\ell-1}$, define a new kernel

$$k_{b_0 b_1 \ldots b_{\ell-1}}(\cdot,\cdot) := k_{b_0 b_1 \ldots b_{\ell-1} 0}(\cdot,\cdot) + k_{b_0 b_1 \ldots b_{\ell-1} 1}(\cdot,\cdot)$$

Repeat this process of labeling with length $\ell - 2$ bitstrings and so on until we have defined k_0 and k_1. Lastly, we define

$$k_\Sigma(\cdot,\cdot) = k_0(\cdot,\cdot) + k_1(\cdot,\cdot) = \sum_{t=1}^m k_t(\cdot,\cdot)$$

.

Now, recall Theorem 2 (or Theorem 6 if we are using the SVM with slack). Let $[b_\ell] := \{b_0 \ldots b_\ell | b \in \{0,1\}\}$ denote the set of all length ℓ bitstrings. Also, for every kernel $k_{b_0\ldots b_j}$, compute the associated kernel matrix $\tilde{\boldsymbol{K}}_{b_0\ldots b_j}$ and dual solution vector $\boldsymbol{\alpha}_{b_0\ldots b_j}$.

Claim 1. *Fix $j \in [\ell - 1]$. Then*

$$\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \leq \left(\frac{1}{3}\right)^j \sum_{b_0\ldots b_j \in [b_j]} \boldsymbol{\alpha}_{b_0\ldots b_j}^\intercal \tilde{\boldsymbol{K}}_{b_0\ldots b_j} \boldsymbol{\alpha}_{b_0\ldots b_j}$$

This claim follows from induction. In the base case, $j = 1$, and Theorem 2 tells us that $\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma \leq \frac{1}{3}(\boldsymbol{\alpha}_0^\intercal \tilde{\boldsymbol{K}}_0 \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1^\intercal \tilde{\boldsymbol{K}}_1 \boldsymbol{\alpha}_1)$, matching the claim. Now, assume the claim holds for $j - 1$. Then,

$$\begin{aligned}\boldsymbol{\alpha}_\Sigma^\intercal \tilde{\boldsymbol{K}}_\Sigma \boldsymbol{\alpha}_\Sigma &\leq \left(\frac{1}{3}\right)^j \sum_{b_0\ldots b_j \in [b_j]} \boldsymbol{\alpha}_{b_0\ldots b_j}^\intercal \tilde{\boldsymbol{K}}_{b_0\ldots b_j} \boldsymbol{\alpha}_{b_0\ldots b_j} \\ &\leq \left(\frac{1}{3}\right)^j \sum_{b_0\ldots b_j \in [b_j]} \frac{1}{3}(\boldsymbol{\alpha}_{b_0\ldots b_j 0}^\intercal \tilde{\boldsymbol{K}}_{b_0\ldots b_j 0} \boldsymbol{\alpha}_{b_0\ldots b_j 0} + \boldsymbol{\alpha}_{b_0\ldots b_j 1}^\intercal \tilde{\boldsymbol{K}}_{b_0\ldots b_j 1} \boldsymbol{\alpha}_{b_0\ldots b_j 1}) \\ &= \left(\frac{1}{3}\right)^{j+1} \sum_{b_0\ldots b_{j+1} \in [b_{j+1}]} \boldsymbol{\alpha}_{b_0\ldots b_{j+1}}^\intercal \tilde{\boldsymbol{K}}_{b_0\ldots b_{j+1}} \boldsymbol{\alpha}_{b_0\ldots b_{j+1}}\end{aligned}$$

This completes the proof of the claim.

Now we need to be careful when moving to the length ℓ kernel labels because if m is not a power of two, then only some of the kernels have a length ℓ label. Let $\mathcal{A}$ be the set of all base kernels that have a length $\ell - 1$ label. Let $\mathcal{B}$ be the rest of the base kernels, with a length ℓ label. By Claim 1, we know that

$$\begin{aligned}
\boldsymbol{\alpha}_{\Sigma}^{\intercal}\tilde{\boldsymbol{K}}_{\Sigma}\boldsymbol{\alpha}_{\Sigma} &\leq \left(\frac{1}{3}\right)^{\ell-1} \sum_{b_0\ldots b_{\ell-1}\in[b_{\ell-1}]} \boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}}^{\intercal}\tilde{\boldsymbol{K}}_{b_0\ldots b_{\ell-1}}\boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}} \\
&= \sum_{b_0\ldots b_{\ell-1}\in[b_{\ell-1}]} \left(\frac{1}{3}\right)^{\ell-1} \boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}}^{\intercal}\tilde{\boldsymbol{K}}_{b_0\ldots b_{\ell-1}}\boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}} \\
&= \sum_{\substack{b_0\ldots b_{\ell-1}\in[b_{\ell-1}]:\\ k_{b_0\ldots b_{\ell-1}}\in\mathcal{A}}} \left(\frac{1}{3}\right)^{\ell-1} \boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}}^{\intercal}\tilde{\boldsymbol{K}}_{b_0\ldots b_{\ell-1}}\boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}} + \sum_{\substack{b_0\ldots b_{\ell-1}\in[b_{\ell-1}]:\\ k_{b_0\ldots b_{\ell-1}}\notin\mathcal{A}}} \left(\frac{1}{3}\right)^{\ell-1} \boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}}^{\intercal}\tilde{\boldsymbol{K}}_{b_0\ldots b_{\ell-1}}\boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}} \\
&\leq \sum_{\substack{b_0\ldots b_{\ell-1}\in[b_{\ell-1}]:\\ k_{b_0\ldots b_{\ell-1}}\in\mathcal{A}}} \left(\frac{1}{3}\right)^{\ell-1} \boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}}^{\intercal}\tilde{\boldsymbol{K}}_{b_0\ldots b_{\ell-1}}\boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}} + \sum_{\substack{b_0\ldots b_{\ell}\in[b_{\ell}]:\\ k_{b_0\ldots b_{\ell}}\in\mathcal{B}}} \left(\frac{1}{3}\right)^{\ell} \boldsymbol{\alpha}_{b_0\ldots b_{\ell}}^{\intercal}\tilde{\boldsymbol{K}}_{b_0\ldots b_{\ell}}\boldsymbol{\alpha}_{b_0\ldots b_{\ell}} \\
&\leq \sum_{\substack{b_0\ldots b_{\ell-1}\in[b_{\ell-1}]:\\ k_{b_0\ldots b_{\ell-1}}\in\mathcal{A}}} \left(\frac{1}{3}\right)^{\ell-1} \boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}}^{\intercal}\tilde{\boldsymbol{K}}_{b_0\ldots b_{\ell-1}}\boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}} + \sum_{\substack{b_0\ldots b_{\ell}\in[b_{\ell}]:\\ k_{b_0\ldots b_{\ell}}\in\mathcal{B}}} \left(\frac{1}{3}\right)^{\ell-1} \boldsymbol{\alpha}_{b_0\ldots b_{\ell}}^{\intercal}\tilde{\boldsymbol{K}}_{b_0\ldots b_{\ell}}\boldsymbol{\alpha}_{b_0\ldots b_{\ell}} \\
&\leq \left(\frac{1}{3}\right)^{\ell-1} \left(\sum_{\substack{b_0\ldots b_{\ell-1}\in[b_{\ell-1}]:\\ k_{b_0\ldots b_{\ell-1}}\in\mathcal{A}}} \boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}}^{\intercal}\tilde{\boldsymbol{K}}_{b_0\ldots b_{\ell-1}}\boldsymbol{\alpha}_{b_0\ldots b_{\ell-1}} + \sum_{\substack{b_0\ldots b_{\ell}\in[b_{\ell}]:\\ k_{b_0\ldots b_{\ell}}\in\mathcal{B}}} \boldsymbol{\alpha}_{b_0\ldots b_{\ell}}^{\intercal}\tilde{\boldsymbol{K}}_{b_0\ldots b_{\ell}}\boldsymbol{\alpha}_{b_0\ldots b_{\ell}} \right) \\
&= \left(\frac{1}{3}\right)^{\ell-1} \sum_{t=1}^{m} \boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t
\end{aligned}$$

Where the second inequalty applies Theorem 2 and the last equality uses the fact that all base kernels are in either $\mathcal{B}$ or $\mathcal{A}$.

Lastly, recall that $\ell = \lceil \log_2(m) \rceil$.

$$\left(\frac{1}{3}\right)^{\ell-1} = 3^{1-\lceil \log_2(m) \rceil} = 3 \cdot 3^{-\lceil \log_2(m) \rceil} \leq 3 \cdot 3^{-\log_2(m)} = 3 \cdot m^{-\log_2(3)}$$

Therefore, overall, we have

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

□

C. Proof of Kernel Sum Rademacher (Theorem 4)

Theorem 4 Restated. *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,*

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

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

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

This proof very closely parallels that of Lemma 22 in (Bartlett & Mendelson, 2002). We produce the entire proof here for completeness. First, note that

$$\mathcal{F}_{\Sigma} \subseteq \{\mathbf{x} \mapsto \mathbf{w}_{\Sigma}^{\intercal}\boldsymbol{\phi}_{\Sigma} \,|\, \|\mathbf{w}_{\Sigma}\|_2 \leq B_{\Sigma}\}$$

Where $\boldsymbol{\phi}_{\Sigma}$ is the concatenation of the feature spaces associated with each of the m kernels, and $B_{\Sigma}^2 = \boldsymbol{\alpha}_{\Sigma}^{\intercal}\tilde{\boldsymbol{K}}_{\Sigma}\boldsymbol{\alpha}_{\Sigma}$. Then,

$$\begin{aligned}
\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}_{\Sigma}) &\leq \frac{1}{n}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}_{\Sigma}\|\leq B_{\Sigma}}\left(\mathbf{w}_{\Sigma}^{\intercal}\sum_{i=1}^{n}\sigma_i y_i \boldsymbol{\phi}_{\Sigma}(\mathbf{x}_i)\right)\right] \\
&= \frac{B_{\Sigma}}{n}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i y_i \boldsymbol{\phi}_{\Sigma}(\mathbf{x}_i)\right\|_2\right] \\
&= \frac{B_{\Sigma}}{n}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sqrt{\left\|\sum_{i=1}^{n}\sigma_i y_i \boldsymbol{\phi}_{\Sigma}(\mathbf{x}_i)\right\|_2^2}\right] \\
&= \frac{B_{\Sigma}}{n}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sqrt{\sum\nolimits_{i,j=1}^{n}\sigma_i\sigma_j[\tilde{\boldsymbol{K}}_{\Sigma}]_{i,j}}\right] \\
&\leq \frac{B_{\Sigma}}{n}\sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum\nolimits_{i,j=1}^{n}\sigma_i\sigma_j[\tilde{\boldsymbol{K}}_{\Sigma}]_{i,j}\right]} \\
&= \frac{B_{\Sigma}}{n}\sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum\nolimits_{i=1}^{n}\sigma_i^2[\tilde{\boldsymbol{K}}_{\Sigma}]_{i,i}\right]} \\
&= \frac{B_{\Sigma}}{n}\sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum\nolimits_{i=1}^{n}[\tilde{\boldsymbol{K}}_{\Sigma}]_{i,i}\right]} \\
&= \frac{B_{\Sigma}}{n}\sqrt{\mathrm{Tr}[\tilde{\boldsymbol{K}}_{\Sigma}]} \\
&= \frac{B_{\Sigma}}{n}\sqrt{\sum_{t=1}^{m}\mathrm{Tr}[\tilde{\boldsymbol{K}}_t]} \\
&\leq \frac{1}{n}\cdot\sqrt{3m^{(1-\log_2(3))}\sum_{t=1}^{m}\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t}\cdot\sqrt{\sum_{t=1}^{m}\mathrm{Tr}[\tilde{\boldsymbol{K}}_t]} \\
&= \frac{1}{n}\sqrt{3m^{-\log_2(3)}\left(\sum_{t=1}^{m}\mathrm{Tr}[\tilde{\boldsymbol{K}}_t]\right)\sum_{t=1}^{m}\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t}
\end{aligned}$$

The second inequality is Jensen's, and the last inequality is Theorem 3. This completes the first part of the proof. We can then substitute in B^2 and R^2:

$$\begin{aligned}
\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}_{\Sigma}) &\leq \frac{1}{n}\sqrt{3m^{-\log_2(3)}\left(\sum_{t=1}^{m}\mathrm{Tr}[\tilde{\boldsymbol{K}}_t]\right)\sum_{t=1}^{m}\boldsymbol{\alpha}_t^{\intercal}\tilde{\boldsymbol{K}}_t\boldsymbol{\alpha}_t} \\
&\leq \frac{1}{n}\sqrt{3m^{-\log_2(3)}\left(\sum_{t=1}^{m}nR^2\right)\sum_{t=1}^{m}B^2} \\
&= \frac{1}{n}\sqrt{3m^{-\log_2(3)}\cdot mnR^2\cdot mB^2} \\
&= \frac{BR}{\sqrt{n}}\sqrt{3m^{(1-\log_2(3/2))}}
\end{aligned}$$

D. Proof of Learning Kernels (Theorem 5)

Theorem 5 Restated. *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,*

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

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

This proof closely follows that of Theorem 1 in (Cortes et al., 2009c).

Proof. Let $s := |\mathcal{P}|$. Let $\mathbf{w}_{\mathcal{P}}$ be the optimal Primal SVM solution using subset of kernels $\mathcal{P}$. Note that $\mathbf{w}_{\mathcal{P}}$ is a *concatenation* of s labeled and scaled feature vectors. To be precise, let $\boldsymbol{\phi}_t$ be the feature map for the t^{th} kernel and define $\mathbf{w}_t := \sum_{i=1}^n \alpha_i y_i \boldsymbol{\phi}_t(\mathbf{x}_i)$. Then $\mathbf{w}_{\mathcal{P}} = [\, \mathbf{w}_{t_1}^\intercal \cdots \mathbf{w}_{t_s}^\intercal \,]^\intercal$, where t_i is the i^{th} smallest element of $\mathcal{P}$.

Consider some $q, r > 1$ such that $\frac{1}{q} + \frac{1}{r} = 1$. Then,

$$\begin{aligned}
\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}_{\mathcal{P}}) &:= \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h\in\mathcal{F}_\Sigma} \sum_{i=1}^n \sigma_i h(\mathbf{x}_i, y_i) \right] \\
&\leq \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{s\in[m]} \sup_{|\mathcal{P}|=s} \sup_{\mathbf{w}_{\mathcal{P}}} \mathbf{w}_{\mathcal{P}}^\intercal \left(\sum\nolimits_{i=1}^n \sigma_i y_i \boldsymbol{\phi}_{\mathcal{P}}(\mathbf{x}_i) \right) \right] \\
&\leq \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{s\in[m]} \sup_{|\mathcal{P}|=s} \sup_{\mathbf{w}_{\mathcal{P}}} \left(\sum\nolimits_{t\in\mathcal{P}} \|\mathbf{w}_t\|_2^q \right)^{1/q} \left(\sum\nolimits_{t\in\mathcal{P}} \left\| \sum\nolimits_{i=1}^n \sigma_i y_i \boldsymbol{\phi}_t(\mathbf{x}_i) \right\|_2^r \right)^{1/r} \right] \\
&\leq \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{s\in[m]} \sup_{|\mathcal{P}|=s} \sup_{\mathbf{w}_{\mathcal{P}}} \left(\sum\nolimits_{t\in\mathcal{P}} \|\mathbf{w}_t\|_2^q \right)^{1/q} \left(\sum\nolimits_{t=1}^m \left\| \sum\nolimits_{i=1}^n \sigma_i y_i \boldsymbol{\phi}_t(\mathbf{x}_i) \right\|_2^r \right)^{1/r} \right] \\
&= \frac{1}{n} \left[\sup_{s\in[m]} \sup_{|\mathcal{P}|=s} \sup_{\mathbf{w}_{\mathcal{P}}} \left(\sum\nolimits_{t=1}^m \|\mathbf{w}_t\|_2^q \right)^{1/q} \right] \cdot \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\left(\sum\nolimits_{t=1}^m \left\| \sum\nolimits_{i=1}^n \sigma_i y_i \boldsymbol{\phi}_t(\mathbf{x}_i) \right\|_2^r \right)^{1/r} \right]
\end{aligned}$$

The third line follows exactly from Lemma 5 in (Cortes et al., 2009c). We bound both terms separately. We only substantially differ from the original proof in bounding the first term. To start, note that $f(x) = x^{1/q}$ is subadditive for $1/q < 1$:

$$\begin{aligned}
\left(\sum\nolimits_{t\in\mathcal{P}} \|\mathbf{w}_t\|_2^q \right)^{1/q} &\leq \sum_{t\in\mathcal{P}} \left(\|\mathbf{w}_t\|_2^q \right)^{1/q} \\
&= \sum_{t\in\mathcal{P}} \left\| \sum_{i=1}^n \alpha_i y_i \boldsymbol{\phi}_t(\mathbf{x}_i) \right\|_2 \\
&= s \sum_{t\in\mathcal{P}} \frac{1}{s} \sqrt{\left\| \sum_{i=1}^n \alpha_i y_i \boldsymbol{\phi}_t(\mathbf{x}_i) \right\|_2^2} \\
&\leq s \sqrt{\sum_{t\in\mathcal{P}} \frac{1}{s} \left\| \sum_{i=1}^n \alpha_i y_i \boldsymbol{\phi}_t(\mathbf{x}_i) \right\|_2^2} \\
&= \sqrt{s \cdot \sum_{t\in\mathcal{P}} \boldsymbol{\alpha}_{\mathcal{P}}^\intercal \tilde{\boldsymbol{K}}_t \boldsymbol{\alpha}_{\mathcal{P}}} \\
&= \sqrt{s \cdot \boldsymbol{\alpha}_{\mathcal{P}}^\intercal \tilde{\boldsymbol{K}}_{\mathcal{P}} \boldsymbol{\alpha}_{\mathcal{P}}} \\
&\leq \sqrt{s \cdot 3 s^{-\log_2(3/2)} B^2} \\
&= B\sqrt{3 s^{(1-\log_2(3/2))}}
\end{aligned}$$

The second inequality follows from Jensen's, and the last inequality is Theorem 2.

We start our bound of the second term by applying Jensen's Inequality:

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left(\sum_{t=1}^{m}\left\|\sum_{i=1}^{n}\sigma_i\boldsymbol{\phi}_t(\mathbf{x}_i)\right\|_2^r\right)^{1/r}\right] &\leq \left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{m}\left\|\sum_{i=1}^{n}\sigma_i\boldsymbol{\phi}_t(\mathbf{x}_i)\right\|_2^r\right]\right)^{1/r} \\ &= \left(\sum_{t=1}^{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i\boldsymbol{\phi}_t(\mathbf{x}_i)\right\|_2^r\right]\right)^{1/r}\end{aligned}$$

We detour to bound the inner expectation. Assume that r is an even integer. That is, $r = 2p$ for some integer p.

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i\boldsymbol{\phi}_t(\mathbf{x}_i)\right\|_2^r\right] &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left(\sum_{i,j=1}^{n}\sigma_i\sigma_j k_t(\mathbf{x}_i,\mathbf{x}_j)\right)^p\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left(\boldsymbol{\sigma}^\intercal\tilde{\boldsymbol{K}}_t\boldsymbol{\sigma}\right)^p\right] \\ &\leq \left(\eta_0 p\operatorname{Tr}[\tilde{\boldsymbol{K}}]\right)^p\end{aligned}$$

Where the last line follows from Lemma 1 in (Cortes et al., 2010), where $\eta_0 = \frac{23}{22}$. Returning to the bound of the second term,

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left(\sum_{t=1}^{m}\left\|\sum_{i=1}^{n}\sigma_i\boldsymbol{\phi}_t(\mathbf{x}_i)\right\|_2^r\right)^{1/r}\right] &\leq \left(\sum_{t=1}^{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i\boldsymbol{\phi}_t(\mathbf{x}_i)\right\|_2^r\right]\right)^{1/2p} \\ &\leq \left(\sum_{t=1}^{m}\left(\eta_0 p\operatorname{Tr}[\tilde{\boldsymbol{K}}_t]\right)^p\right)^{1/2p} \\ &\leq \left(\sum_{t=1}^{m}\left(\eta_0 p\, nR^2\right)^p\right)^{1/2p} \\ &= \left(m\left(\eta_0 p\, nR^2\right)^p\right)^{1/2p} \\ &= m^{1/2p}\sqrt{\eta_0 p\, nR^2}\end{aligned}$$

By differentiating, we find that $p = \ln(m)$ minimizes this expression. We required p to be an integer, so we instead take $p = \lceil\ln(m)\rceil$.

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left(\sum_{t=1}^{m}\left\|\sum_{i=1}^{n}\sigma_i\boldsymbol{\phi}_t(\mathbf{x}_i)\right\|_2^r\right)^{1/r}\right] &\leq Rm^{1/2p}\sqrt{\eta_0 pn} \\ &= Rm^{\frac{1}{2\lceil\ln(m)\rceil}}\sqrt{\eta_0\lceil\ln(m)\rceil\, n} \\ &\leq R\sqrt{e\eta_0\lceil\ln(m)\rceil\, n}\end{aligned}$$

Combining the first and second terms' bounds, we return to the bound of the Rademacher complexity itself:

$$\begin{aligned}\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}_{\mathcal{P}}) &\leq \frac{1}{n}\left[\sup_{s\in[m]}\sup_{|\mathcal{P}|=s}\sup_{\mathbf{w}_{\mathcal{P}}}\left(\sum_{t=1}^{m}\|\mathbf{w}_t\|_2^q\right)^{1/q}\right]\cdot\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left(\sum_{t=1}^{m}\left\|\sum_{i=1}^{n}\sigma_i y_i\boldsymbol{\phi}_t(\mathbf{x}_i)\right\|_2^r\right)^{1/r}\right] \\ &\leq \frac{1}{n}\left[\sup_{s\in[m]} B\sqrt{3s^{(1-\log_2(3/2))}}\right]\cdot\left[R\sqrt{e\eta_0\lceil\ln(m)\rceil\, n}\right] \\ &= \frac{1}{n}\left[B\sqrt{3m^{(1-\log_2(3/2))}}\right]\cdot\left[R\sqrt{e\eta_0\lceil\ln(m)\rceil\, n}\right] \\ &= \frac{BR\sqrt{3e\eta_0\, m^{(1-\log_2(3/2))}\lceil\ln(m)\rceil}}{\sqrt{n}}\end{aligned}$$

□