A. Proof of theorem 1

Proof. Let $\widetilde{H}$ be the family of hypotheses mapping $\mathcal{X}\times\mathcal{Y}$ to $\mathbb{R}$ defined by $\widetilde{H} = \{z = (x,y) \mapsto \rho_h(x,y) \colon h \in H\}$. Consider the family of functions $\widetilde{\mathcal{H}} = \{\Phi_\rho \circ r \colon r \in \widetilde{H}\}$ derived from $\widetilde{H}$, where Φ_ρ is the ρ-margin function loss defined by $\Phi_\rho(x) = 1_{x\leq 0} + \max(0, 1 - x/\rho)\, 1_{x>0}$. By the Rademacher complexity bound for functions taking values in $[0,1]$ (see (Koltchinskii & Panchenko, 2002; Bartlett & Mendelson, 2002)), we can write that with probability at least $1-\delta$, for all $h \in H$,

$$\mathrm{E}\left[\Phi_\rho(\rho_h(x,y))\right] \leq \widehat{R}_\rho(h) + 2\widehat{\mathfrak{R}}_S\big(\Phi_\rho \circ \widetilde{H}\big) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\,.$$

Since $1_{u\leq 0} \leq \Phi_\rho(u)$ for all $u \in \mathbb{R}$, the generalization error $R(h)$ is a lower bound on the left-hand side, $R(h) = \mathrm{E}[1_{y[h(x')-h(x)]\leq 0}] \leq \mathrm{E}\left[\Phi_\rho(\rho_h(x,y))\right]$. Furthermore, since Φ_ρ is $1/\rho$-Lipschitz, by Talagrand's contraction lemma (Ledoux & Talagrand, 1991), we have $\widehat{\mathfrak{R}}_S\big(\Phi_\rho \circ \widetilde{H}\big) \leq \frac{1}{\rho}\widehat{\mathfrak{R}}_S(\widetilde{H})$.

For any fixed $y \in \mathcal{Y}$ and any $i \in [1,m]$, define ϵ_i as $2(1_{y=y_i}) - 1$. Since $\epsilon_i \in \{-1, +1\}$, the random variables σ_i and $\sigma_i\epsilon_i$ follow the same distribution. Using this fact and the sub-additivity of sup, $\widehat{\mathfrak{R}}_S(\widetilde{H})$ can be upper bounded as follows:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\widetilde{H}) &= \frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\sum_{i=1}^m \sigma_i\rho_h(x_i,y_i)\Big]\\
&= \frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\sum_{i=1}^m\sum_{y\in\mathcal{Y}} \sigma_i\rho_h(x_i,y)1_{y=y_i}\Big]\\
&\leq \frac{1}{m}\sum_{y\in\mathcal{Y}}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\sum_{i=1}^m \sigma_i\rho_h(x_i,y)1_{y=y_i}\Big]\\
&= \frac{1}{m}\sum_{y\in\mathcal{Y}}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\sum_{i=1}^m \sigma_i\rho_h(x_i,y)\big(\tfrac{2(1_{y=y_i})-1}{2} + \tfrac{1}{2}\big)\Big]\\
&\leq \frac{1}{2m}\sum_{y\in\mathcal{Y}}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\sum_{i=1}^m \sigma_i\epsilon_i\rho_h(x_i,y)\Big]\\
&\quad + \frac{1}{2m}\sum_{y\in\mathcal{Y}}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\sum_{i=1}^m \sigma_i\rho_h(x_i,y)\Big]\\
&= \frac{1}{m}\sum_{y\in\mathcal{Y}}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\sum_{i=1}^m \sigma_i\rho_h(x_i,y)\Big].
\end{aligned}$$

Let $H_{\mathcal{X}}^{(c-1)} = \big\{\max\{h_1,\ldots,h_l\} \colon h_i \in H_{\mathcal{X}}, i \in [1, c-1]\big\}$. It is known that the empirical Rademacher of a function class defined as that of the maxima of several hypotheses is upper bounded by the sum of the empirical Rademacher complexities of the sets to which each of these hypotheses belong to (Ledoux & Talagrand, 1991), thus $\widehat{\mathfrak{R}}_S(H_{\mathcal{X}}^{(c-1)}) \leq (c-1)\widehat{\mathfrak{R}}_S(H_{\mathcal{X}})$.

Now, rewriting $\rho_h(x_i,y)$ explicitly, using again the sub-additivity of sup and observing that $-\sigma_i$ and σ_i are distributed in the same way leads to

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\widetilde{H}) &\leq \frac{1}{m}\sum_{y\in\mathcal{Y}}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\sum_{i=1}^m \sigma_i\big(h(x_i,y) - \max_{y'\neq y} h(x_i,y')\big)\Big]\\
&\leq \sum_{y\in\mathcal{Y}}\Bigg[\frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\sum_{i=1}^m \sigma_i h(x_i,y)\Big]\\
&\qquad + \frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\sum_{i=1}^m -\sigma_i \max_{y'\neq y} h(x_i,y')\Big]\Bigg]\\
&= \sum_{y\in\mathcal{Y}}\Bigg[\frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\sum_{i=1}^m \sigma_i h(x_i,y)\Big]\\
&\qquad + \frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H}\sum_{i=1}^m \sigma_i \max_{y'\neq y} h(x_i,y')\Big]\Bigg]\\
&\leq \sum_{y\in\mathcal{Y}}\Bigg[\frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H_{\mathcal{X}}}\sum_{i=1}^m \sigma_i h(x_i)\Big]\\
&\qquad + \frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H_{\mathcal{X}}^{(c-1)}}\sum_{i=1}^m \sigma_i h(x_i)\Big]\Bigg]\\
&\leq c\left[\frac{c}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h\in H_{\mathcal{X}}}\sum_{i=1}^m \sigma_i h(x_i)\Big]\right] = c^2\widehat{\mathfrak{R}}_S(H_{\mathcal{X}}).
\end{aligned}$$

This concludes the proof. □

B. Proof of lemma 1

Proof. For any $h \in \mathbb{H}_K$ and $x \in \mathcal{X}$, by the reproducing property, we have $h(x) = \langle h, K(x,\cdot)\rangle$. Let $\mathbb{H}_S = \mathrm{span}(\{K(x_i,\cdot) \colon i \in [1,m]\})$, then, for $i \in [1,m]$, $h(x_i) = \langle h', K(x,\cdot)\rangle$, where h' is the orthogonal projection of h over $\mathbb{H}_S$. Thus, there exists $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_m)^\top \in \mathbb{R}^m$ such that $h' = \sum_{i=1}^m \alpha_i K(x_i,\cdot)$. If $\|h\|_{\mathbb{H}} \leq \Lambda$, then $\boldsymbol{\alpha}^\top\mathbf{K}\boldsymbol{\alpha} = \|h'\|_K^2 \leq \|h\|_K^2 \leq \Lambda^2$ where $\mathbf{K}$ is the kernel matrix of K for the sample S. Conversely, any $\sum_{i=1}^p \alpha_i K(x_i,\cdot)$ with $\boldsymbol{\alpha}^\top\mathbf{K}\boldsymbol{\alpha} \leq \Lambda^2$ is the projection of some $h \in \mathbb{H}_K$ with $\|h\|_K^2 \leq \Lambda^2$.

Thus, for any $y \in \mathcal{Y}$, there exists $\boldsymbol{\alpha}^y = (\alpha_1^y,\ldots,\alpha_m^y)^\top \in \mathbb{R}^m$ such that for any $i \in [1,m]$, $h_y(x_i) = \sum_{j=1}^m \alpha_j^y K_{\boldsymbol{\mu}}(x_i,x_j)$, and $\boldsymbol{\alpha}^{y\top}\mathbf{K}_{\boldsymbol{\mu}}\boldsymbol{\alpha}^y \leq \Lambda$ where $\mathbf{K}_{\boldsymbol{\mu}}$ is the kernel matrix associated to $K_{\boldsymbol{\mu}}$ for the sample $(x_1,\ldots,x_m)$. In view of that, we can write

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(H_{\mathcal{X}}^1) &= \frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{h_y\in H_{\mathcal{X}}^1}\sum_{i=1}^m \sigma_i h_y(x_i)\Big]\\
&= \frac{1}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{\boldsymbol{\mu}\in\mathcal{M}_1, y\in\mathcal{Y}, \boldsymbol{\alpha}^y} \boldsymbol{\sigma}\mathbf{K}_{\boldsymbol{\mu}}\boldsymbol{\alpha}^y\Big].
\end{aligned}$$

Now, by the Cauchy-Schwarz inequality, the supremum $\sup_{\boldsymbol{\alpha}^y} \boldsymbol{\sigma}^\top\mathbf{K}_{\boldsymbol{\mu}}\boldsymbol{\alpha}^y$ is reached for $\mathbf{K}_{\boldsymbol{\mu}}^{1/2}\boldsymbol{\alpha}^y$ collinear

with $\mathbf{K}_{\boldsymbol{\mu}}^{1/2}\boldsymbol{\sigma}$, which gives $\sup_{\boldsymbol{\alpha}^y \in \mathcal{A}} \boldsymbol{\sigma}^\top \mathbf{K}_{\boldsymbol{\mu}} \boldsymbol{\alpha}^y = \Lambda\sqrt{\boldsymbol{\sigma}^\top \mathbf{K}_{\boldsymbol{\mu}} \boldsymbol{\sigma}}$. Thus,

$$\begin{aligned}\widehat{\mathfrak{R}}_S(H^1_{\mathcal{X}}) &= \frac{\Lambda}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{\boldsymbol{\mu}\in\mathcal{M}_1} \sqrt{\boldsymbol{\sigma}^\top \mathbf{K}_{\boldsymbol{\mu}} \boldsymbol{\sigma}}\Big] \\ &= \frac{\Lambda}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\Big[\sup_{\boldsymbol{\mu}\in\mathcal{M}_1} \sqrt{\boldsymbol{\mu}\cdot \mathbf{u}_{\boldsymbol{\sigma}}}\Big],\end{aligned}$$

which concludes the proof. □

C. Proof of lemma 3

Proof. It first helps to rewrite the optimization in the following equivalent form:

$$\min_{\lambda\geq 0, t} \; t \quad \text{subject to: } \forall k \in [1,p],\; t \geq \mathbf{u}_{\boldsymbol{\sigma},k} + \lambda(\widetilde{\gamma}_k - \gamma_0).$$

The Lagrangian L associated to this problem can be defined for any $t \in \mathbb{R}$, $\lambda \geq 0$, and $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\beta} \geq 0$, by

$$L(t,\lambda,\boldsymbol{\beta}) = t + \sum_{k=1}^p \beta_k\big(\mathbf{u}_{\boldsymbol{\sigma},k} + \lambda(\widetilde{\gamma}_k - \gamma_0) - t\big).$$

By the KKT conditions, the following holds:

$$\frac{\partial L}{\partial t} = 1 - \sum_{k=1}^p \beta_k^* = 0 \iff \sum_{k=1}^p \beta_k^* = 1, \tag{10}$$

$$\frac{\partial L}{\partial \lambda} = \sum_{k=1}^p \beta_k^*(\widetilde{\gamma}_k - \gamma_0) = 0, \tag{11}$$

$$\forall k : \beta_k^* > 0,\; \mathbf{u}_{\boldsymbol{\sigma},k} + \lambda^*(\widetilde{\gamma}_k - \gamma_0) = t^*. \tag{12}$$

Here λ^*, t^*, and $\boldsymbol{\beta}^*$ denote the optimal values for the dual problem $\max_{\boldsymbol{\beta}\geq 0} \min_{\lambda,t} L(t,\lambda,\boldsymbol{\beta})$. Note that Slater's condition holds for the convex primal problem, which implies that strong duality holds and that λ^* and t^* are also optimal solutions to the primal problem. The conditions (10) and (11) follow from the fact that $\nabla L = \mathbf{0}$ at the optimum and (12) follows from the complementary slackness condition that guarantees for all k, $\beta_k^*(\mathbf{u}_{\boldsymbol{\sigma},k} + \lambda^*(\widetilde{\gamma}_k - \gamma_0) - t^*) = 0$. Note that the condition in (10) as well as the constraint $\beta_k \geq 0$, which is imposed on dual variables that correspond to inequality constraints, implies $\boldsymbol{\beta}^* \in \Delta_1$.

We first consider the subset of solutions to the dual optimization that are equal to $T_1 = \max_{k:\widetilde{\gamma}_k \geq \gamma_0} \mathbf{u}_{\boldsymbol{\sigma},k}$. Note that any feasible point in this simpler optimization is also a feasible point in the original problem. Now, consider the dual optimization, which can be simplified by removing t when we impose $\boldsymbol{\beta} \in \Delta_1$: $\max_{\boldsymbol{\beta}\in\Delta_1} \min_\lambda \sum_{k=1}^p \beta_k(\mathbf{u}_{\boldsymbol{\sigma},k} + \lambda(\widetilde{\gamma}_k - \gamma_0))$. Whenever it is the case that $\lambda^* = 0$ or that $\widetilde{\gamma}_k = \gamma_0$ for all k where $\beta_k > 0$, we can further simplify the objective to $\max_{\boldsymbol{\beta}\in\Delta_1} \sum_{k=1}^p \beta_k \mathbf{u}_{\boldsymbol{\sigma},k} = \max_k \mathbf{u}_{\boldsymbol{\sigma},k}$. Note this implies that $\lambda^* = 0$ iff $\widetilde{\gamma}_{k_{\max}} \geq \gamma_0$ (where $k_{\max} = \operatorname{argmax}_k \mathbf{u}_{\boldsymbol{\sigma},k}$), since all constraints of the original problem in lemma 2 must be satisfied at the optimum. Thus, T_1 is found as the solution to the dual problem whenever $\lambda^* = 0$ or $\widetilde{\gamma}_k = \gamma_0$ for all k where $\beta_k > 0$.

Now we seek an expression T_2 that is equal to the optimum of the dual optimization in the cases not accounted for by T_1. That is, we consider the case $\lambda^* > 0$ and the existence of at least one k such that $\beta_k^* > 0$ and $\widetilde{\gamma}_k \neq \gamma_0$. In order for condition (11) to be satisfied in this case, there must be at least two coordinates β_k^* and $\beta_{k'}^*$ that are non-zero for k and k' that satisfy $\widetilde{\gamma}_k < \gamma_0 < \widetilde{\gamma}_{k'}$. This is because both a negative coefficient, i.e. $(\widetilde{\gamma}_k - \gamma_0)$, as well as a positive coefficient, i.e. $(\widetilde{\gamma}_k - \gamma_0)$, must be present in order for 0 to be found as convex combination. Now, fix any two coordinates k and k' that are non-zero in $\boldsymbol{\beta}^*$ and that satisfy $\widetilde{\gamma}_k \leq \gamma_0 \leq \widetilde{\gamma}_{k'}$ with $\widetilde{\gamma}_k \neq \widetilde{\gamma}_{k'}$ (there exists *at least* two such coordinates by the argument just discussed). From condition (12) we know that

$$\begin{aligned}&\mathbf{u}_{\boldsymbol{\sigma},k} + \lambda^*(\widetilde{\gamma}_k - \gamma_0) = \mathbf{u}_{\boldsymbol{\sigma},k'} + \lambda^*(\widetilde{\gamma}_{k'} - \gamma_0) \\ &\iff \lambda^* = \frac{\mathbf{u}_{\boldsymbol{\sigma},k} - \mathbf{u}_{\boldsymbol{\sigma},k'}}{\widetilde{\gamma}_k - \widetilde{\gamma}_{k'}} > 0.\end{aligned} \tag{13}$$

Plugging this back into (12) we find an expression for the optimal objective value t^*:

$$t^* = \alpha_{k,k'}\mathbf{u}_{\boldsymbol{\sigma},k} + (1-\alpha_{k,k'})\mathbf{u}_{\boldsymbol{\sigma},k'},$$

where $0 \leq \alpha_{k,k'} = \frac{\widetilde{\gamma}_{k'} - \gamma_0}{\widetilde{\gamma}_{k'} - \widetilde{\gamma}_k} \leq 1$. However, we still do not know which k and k' are active at the optimum. We do know that for all k, $t^* \geq \mathbf{u}_{\boldsymbol{\sigma},k} + \lambda^*(\widetilde{\gamma}_k - \gamma_0)$ since all constraints must hold at the optimum, which also implies, for all k and k',

$$\begin{aligned}t^* &\geq \alpha_{k,k'}(\mathbf{u}_{\boldsymbol{\sigma},k} + \lambda^*(\widetilde{\gamma}_k - \gamma_0)) \\ &\quad + (1-\alpha_{k,k'})(\mathbf{u}_{\boldsymbol{\sigma},k'} + \lambda^*(\widetilde{\gamma}_{k'} - \gamma_0)) \\ &= \alpha_{k,k'}\mathbf{u}_{\boldsymbol{\sigma},k} + (1-\alpha_{k,k'})\mathbf{u}_{\boldsymbol{\sigma},k'} \\ &\quad + \lambda^* \underbrace{\big(\alpha_{k,k'}(\widetilde{\gamma}_k - \gamma_0)) + (1-\alpha_{k,k'})(\widetilde{\gamma}_{k'} - \gamma_0)\big)}_{=\,0}.\end{aligned}$$

Thus, we can maximize over all feasible choice of k and k' in order to find the value at which the above inequality is tight:

$$t^* = \max_{(k,k')\in J_p} \alpha_{k,k'}\mathbf{u}_{\boldsymbol{\sigma},k} + (1-\alpha_{k,k'})\mathbf{u}_{\boldsymbol{\sigma},k'},$$

which gives us the expression T_2 for the optimum in the intersection case. Finally, taking the maximum over T_1 and T_2 completes the lemma. □

D. Theorem 3

Theorem 3. *Fix $\rho > 0$ and let $p' = \mathrm{Card}(I_p) \leq p$ and $p'' = \mathrm{Card}(J_p) < p^2$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over a sample of size m, the following multi-class classification generalization bound holds for all $h \in H^1$:*

$$R(h) \leq \widehat{R}_\rho(h) + \frac{2c^2\Lambda}{m\rho}\sqrt{T_{\gamma_0} + m\lambda_{\max}\sqrt{\frac{\log(p'+p'')}{2}}} + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}, \quad (14)$$

where $T_{\gamma_0} = \max(\max_{k\in I_p}\mathrm{Tr}[\mathbf{K}_k], \max_{(k,k')\in J_p}\mathrm{Tr}[\mathbf{K}_{k,k'}])$ *and* $\lambda_{\max} = \max\left(\max_{k\in I_p}\|\mathbf{K}_k\|_2, \max_{(k,k')\in J_p}\|\mathbf{K}_{k,k'}\|_2\right)$, *with* $\mathbf{K}_{k,k'} = \alpha_{k,k'}\mathbf{K}_k + (1-\alpha_{k,k'})\mathbf{K}_{k'}$ *and* $\alpha_{k,k'} = \frac{\widetilde{\gamma}_{k'}-\gamma_0}{\widetilde{\gamma}_{k'}-\widetilde{\gamma}_k}$.

Proof. Let $M_p = \left\{(k,k') \in [0,p]\times[1,p] \colon (\widetilde{\gamma}_k \leq \gamma_0 \leq \widetilde{\gamma}_{k'}) \wedge (\widetilde{\gamma}_k \neq \widetilde{\gamma}_{k'})\right\}$ as in the proof of Theorem 2. By lemmas 1-3 and Jensen's inequality, we can write:

$$\widehat{\mathfrak{R}}_S(H^1_{\mathcal{X}}) \leq \frac{\Lambda}{m}\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sqrt{\max_{(k,k')\in M_p}\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma}}\right] \leq \frac{\Lambda}{m}\sqrt{\mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\max_{(k,k')\in M_p}\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma}\right]}.$$

Note that for any k, k', $\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma} \leq \|\boldsymbol{\sigma}\|^2\|\mathbf{K}_{k,k'}\|_2 = m\|\mathbf{K}_{k,k'}\|_2$. Now, for any $t \in \mathbb{R}$, by the convexity of exp and Jensen's inequality, we have

$$\begin{aligned} e^{t\,\mathrm{E}[\max_{(k,k')\in M_p}\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma}]} &\leq \mathrm{E}[e^{t\max_{(k,k')\in M_p}\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma}}] \\ &= \mathrm{E}[\max_{(k,k')\in M_p} e^{t\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma}}] \\ &\leq \mathrm{E}[\sum_{(k,k')\in M_p} e^{t\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma}}]. \end{aligned}$$

For any $(k,k') \in M_p$, we have $\mathrm{E}[\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma}] = \mathrm{Tr}[\mathbf{K}_{k,k'}]$. Thus, by Hoeffding's inequality, the following holds

$$\begin{aligned} \mathrm{E}[e^{t\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma}}] &= e^{t\,\mathrm{Tr}[\mathbf{K}_{k,k'}]}\,\mathrm{E}[e^{t(\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma} - \mathrm{E}[\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma}])}] \\ &\leq e^{t\,\mathrm{Tr}[\mathbf{K}_{k,k'}]}e^{t^2\lambda^2_{\max}m^2/8}. \end{aligned}$$

Therefore, we can write

$$e^{t\,\mathrm{E}[\max_{(k,k')\in M_p}\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma}]} \leq (p'+p'')e^{t\max_{k,k'}\mathrm{Tr}[\mathbf{K}_{k,k'}]}e^{t^2\lambda^2_{\max}m^2/8}.$$

Taking the log of both sides and rearranging gives

$$\mathrm{E}[\max_{(k,k')\in M_p}\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma}] \leq \max_{(k,k')\in M_p}\mathrm{Tr}[\mathbf{K}_{k,k'}] + \frac{\log(p'+p'')}{t} + t\lambda^2_{\max}m^2/8.$$

Choosing $t = \sqrt{8(\log(p'+p''))/(\lambda^2_{\max}m^2)}$ to minimize the upper bound gives

$$\mathrm{E}[\max_{(k,k')\in M_p}\boldsymbol{\sigma}^\top\mathbf{K}_{k,k'}\boldsymbol{\sigma}] \leq \max_{(k,k')\in M_p}\mathrm{Tr}[\mathbf{K}_{k,k'}] + m\lambda_{\max}\sqrt{\frac{\log(p'+p'')}{2}}.$$

Plugging in this upper bound on the Rademacher complexity of $H^1_{\mathcal{X}}$ in the learning guarantee of theorem 1 concludes the proof. □

E. Proof of lemma 4

Proof. We introduce $M_1(\gamma_0) = \max_{k\in I_p(\gamma_0)}\mathrm{Tr}[\mathbf{K}_k]$ and $M'_1(\gamma_0) = \max_{(k,k')\in J_p(\gamma_0)}\mathrm{Tr}[\mathbf{K}_{k,k'}]$, where we write $I_p(\gamma_0)$ and $J_p(\gamma_0)$ to make the dependency of these sets on γ_0 explicit. With these definitions, we can write $T_{\gamma_0} = \max(M_1(\gamma_0), M'_1(\gamma_0))$.

Notice that if $\mathrm{Tr}[K_k] \leq \mathrm{Tr}[K_{k'}]$, then $\mathrm{Tr}[K_{kk'}] \leq \mathrm{Tr}[K_{k'}]$ since $\mathrm{Tr}[K_{kk'}]$ is a convex combination of $\mathrm{Tr}[K_k]$ and $\mathrm{Tr}[K_{k'}]$. Thus, if (k,k') is the maximizing pair of indices defining $M'_1(\gamma_0)$ and $\mathrm{Tr}[K_k] \leq \mathrm{Tr}[K_{k'}]$, we have $M'_1(\gamma_0) = \mathrm{Tr}[K_{kk'}] \leq \mathrm{Tr}[K_{k'}] \leq \max_{\gamma_{k'}\geq\gamma_0}\mathrm{Tr}[K_{k'}] = M_1(\gamma_0)$. By contraposition, if $M_1(\gamma_0) < M'_1(\gamma_0)$, then $\mathrm{Tr}[K_{k'}] < \mathrm{Tr}[K_k]$. In view of that, we can rewrite $T_{\gamma_0} = \max(M_1(\gamma_0), M_2(\gamma_0))$, where $M_2(\gamma_0) = \max_{(k,k')\in N_p(\gamma_0)}\mathrm{Tr}[\mathbf{K}_{k,k'}]$, where $N_p(\gamma_0) = \left\{(k,k') \in [1,p]^2 \colon (\widetilde{\gamma}_k \leq \gamma_0 \leq \widetilde{\gamma}_{k'}) \wedge (\widetilde{\gamma}_k \neq \widetilde{\gamma}_{k'}) \wedge (\mathrm{Tr}[K_{k'}] < \mathrm{Tr}[K_k])\right\}$.

Now, let $\lambda_0 \geq \gamma_0$, we will show that $T_{\lambda_0} \leq T_{\gamma_0}$. First note that if $T_{\lambda_0} = M_1(\lambda_0)$, then, since by definition of M_1, $M_1(\lambda_0) \leq M_1(\gamma_0) \leq T_{\gamma_0}$, this shows immediately that $T_{\lambda_0} \leq T_{\gamma_0}$.

Otherwise, $T_{\lambda_0} = M_2(\lambda_0)$. Let (l,l') be the maximizing indices in the definition of $M_2(\lambda_0)$. Then, since $\mathrm{Tr}[K_{l'}] - \mathrm{Tr}[K_l] < 0$ holds, we can write:

$$\begin{aligned} T_{\lambda_0} &= \frac{\widetilde{\gamma}_{l'}-\lambda_0}{\widetilde{\gamma}_{l'}-\widetilde{\gamma}_l}\mathrm{Tr}[K_l] + \frac{\lambda_0-\widetilde{\gamma}_l}{\widetilde{\gamma}_{l'}-\widetilde{\gamma}_l}\mathrm{Tr}[K_{l'}] \\ &= \frac{\widetilde{\gamma}_{l'}\,\mathrm{Tr}[K_l] - \widetilde{\gamma}_l\,\mathrm{Tr}[K_{l'}]}{\widetilde{\gamma}_{l'}-\widetilde{\gamma}_l} + \frac{\lambda_0(\mathrm{Tr}[K_{l'}] - \mathrm{Tr}[K_l])}{\widetilde{\gamma}_{l'}-\widetilde{\gamma}_l} \\ &\leq \frac{\widetilde{\gamma}_{l'}\,\mathrm{Tr}[K_l] - \widetilde{\gamma}_l\,\mathrm{Tr}[K_{l'}]}{\widetilde{\gamma}_{l'}-\widetilde{\gamma}_l} + \frac{\gamma_0(\mathrm{Tr}[K_{l'}] - \mathrm{Tr}[K_l])}{\widetilde{\gamma}_{l'}-\widetilde{\gamma}_l} \\ &= \frac{\widetilde{\gamma}_{l'}-\gamma_0}{\widetilde{\gamma}_{l'}-\widetilde{\gamma}_l}\mathrm{Tr}[K_l] + \frac{\gamma_0-\widetilde{\gamma}_l}{\widetilde{\gamma}_{l'}-\widetilde{\gamma}_l}\mathrm{Tr}[K_{l'}]. \end{aligned}$$

If $(l,l') \in N_p(\gamma_0)$, the right-hand side is upper bounded by $M_2(\gamma_0) \leq T_{\gamma_0}$. Otherwise, if (l,l') is not in $N_p(\gamma_0)$, since by definition of $N_p(\gamma_0)$, $\widetilde{\gamma}_{l'} \geq \lambda_0 \geq \gamma_0$, this can only be because $\gamma_0 \leq \widetilde{\gamma}_l$. But in that case both $\widetilde{\gamma}_l$ and $\widetilde{\gamma}_{l'}$ are greater than or equal to γ_0. Then, by definition of $M_1(\gamma_0)$, $\mathrm{Tr}[K_{ll'}] \leq \max(\mathrm{Tr}[K_l], \mathrm{Tr}[K_{l'}]) \leq$

$M_1(\gamma_0) \leq T_{\gamma_1}$. Thus, the inequality $T_{\lambda_0} \leq T_{\gamma_0}$ holds in all cases. □

F. Alternative M^{3}K optimization

Here we present an alternative formulation of the M^{3}K algorithm, which results in a quadratically constrained linear program. Such a problem can be solved with any standard second order cone programming (SOCP) solver. We find this formulation can be faster to solve than the SILP formulation for smaller size problems, especially in the case of fewer classes.

First, consider the dual formulation with the margin constraint appearing instead as an additional penalty term in the objective with regularization parameter Γ. For every choice of γ_0 in the constraint version of the optimization there exists a choice of Γ that results in an equivalent optimization problem.

$$\begin{aligned}\min_{\boldsymbol{\mu},\boldsymbol{\gamma}} \max_{\boldsymbol{\alpha}} \; & \sum_{i=1}^{m} \boldsymbol{\alpha}_i^\top \mathbf{e}_{y_i} - \frac{C}{2} \sum_{i,j=1}^{m} \boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_j \sum_{k=1}^{p} \mu_k K_k(x_i, x_j) \\ & - \Gamma \sum_{i=1}^{m} \gamma_i \\ \text{s.t. } & \forall i,\ \boldsymbol{\alpha}_i \leq \mathbf{e}_{y_i},\ \boldsymbol{\alpha}_i^\top \mathbf{1} = 0 \\ & \forall i, \forall y \neq y_i,\ \gamma_i \leq \boldsymbol{\mu}^\top \boldsymbol{\eta}(x_i, y_i, y) \\ & \boldsymbol{\mu} \geq \mathbf{0},\ \boldsymbol{\mu}^\top \mathbf{1} = \Lambda\,.\end{aligned}$$

The objective is linear in $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ and concave in $\boldsymbol{\alpha}$ and both $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ are drawn from convex compact sets. The $\boldsymbol{\gamma}$ is drawn from a closed convex set, however, it is unbounded from below. One can add a lower bound on $\boldsymbol{\gamma}$ that has no effect on the optimal solution in order to achieve compactness and, thus, the minimax theorem applies and we permute the min and the max and solve for $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ first. Focusing on the terms that depend on $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$, we have:

$$\begin{aligned}\min_{\boldsymbol{\mu},\boldsymbol{\gamma}} \; & -\frac{C}{2} \boldsymbol{\mu}^\top \mathbf{u} - \Gamma \sum_{i=1}^{m} \gamma_i \\ \text{s.t. } & \forall i, \forall y \neq y_i,\ \gamma_i \leq \boldsymbol{\mu}^\top \boldsymbol{\eta}(x_i, y_i, y) \\ & \boldsymbol{\mu} \geq \mathbf{0},\ \boldsymbol{\mu}^\top \mathbf{1} = \Lambda\,,\end{aligned}$$

where $u_k = \sum_{i,j=1}^{m} \boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_j K_k(x_i, x_j)$. Consider the partial Lagrangian for the constraints on $\boldsymbol{\gamma}$, which introduces dual variables $\boldsymbol{\beta} \geq \mathbf{0}$:

$$\begin{aligned}L = & -\frac{C}{2} \boldsymbol{\mu}^\top \mathbf{u} - \Gamma \sum_{i=1}^{m} \gamma_i \\ & + \sum_{i=1}^{m} \sum_{y \neq y_i} \beta_{i,y} (\gamma_i - \boldsymbol{\mu}^\top \boldsymbol{\eta}(x_i, y_i, y))\,.\end{aligned}$$

At the optimum, it is guaranteed that

$$\forall i,\ \frac{\partial L}{\partial \gamma_i} = -\Gamma + \sum_{y \neq y_i} \beta_{i,y} = 0 \iff \sum_{y \neq y_i} \beta_{i,y} = \Gamma\,.$$

Enforcing this constraint explicitly gives the following simplified objective, where $\boldsymbol{\gamma}$ is no longer appears (since $\Gamma \sum_i \gamma_i - \sum_{i,y \neq y_i} \beta_{i,y} \gamma_i = 0$):

$$\begin{aligned}\min_{\boldsymbol{\mu}} \max_{\boldsymbol{\beta} \geq 0} \; & -\frac{C}{2} \boldsymbol{\mu}^\top \mathbf{u} - \sum_{i=1}^{m} \sum_{y \neq y_i} \boldsymbol{\mu}^\top \boldsymbol{\beta}_{i,y} \boldsymbol{\eta}(x_i, y_i, y)) \\ \text{s.t. } & \forall i,\ \sum_{y \neq y_i} \beta_{i,y} = \Gamma \\ & \boldsymbol{\mu} \geq \mathbf{0},\ \boldsymbol{\mu}^\top \mathbf{1} = \Lambda\,.\end{aligned}$$

Note the minimax theorem applies once again (the objective is linear in $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ which are both drawn from convex compact sets) and we can first solve the minimization over $\boldsymbol{\mu}$ for a fixed $\boldsymbol{\beta}$. Since the objective and the constraints are linear in $\boldsymbol{\mu}$ it is sufficient to consider a solution that places all of the possible mass on a single coordinate of $\boldsymbol{\mu}$. Thus, an equivalent objective is: $\min_{\boldsymbol{\mu}} \boldsymbol{\mu}^\top (-\frac{C}{2}\mathbf{u} - \sum_{i=1}^{m} \sum_{y \neq y_i} \boldsymbol{\beta}_{i,y} \boldsymbol{\eta}(x_i, y_i, y)) = \min_k \Lambda(-\frac{C}{2} u_k - \sum_{i=1}^{m} \sum_{y \neq y_i} \boldsymbol{\beta}_{i,y} \eta_k(x_i, y_i, y))$. A final reformulation, via the introduction of the variable t, simplifies the objective further:

$$\begin{aligned}\max_{\boldsymbol{\beta}, t} \; & \Lambda t \\ \text{s.t. } & \forall k,\ t \leq -\frac{C}{2} u_k - \sum_{i=1}^{m} \sum_{y \neq y_i} \boldsymbol{\beta}_{i,y} \eta_k(x_i, y_i, y) \\ & \forall i,\ \sum_{y \neq y_i} \beta_{i,y} = \Gamma,\ \boldsymbol{\beta} \geq \mathbf{0}\,.\end{aligned}$$

Plugging this solution for $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ back into the original optimization gives the final overall quadratically constraint linear program:

$$\begin{aligned}\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}, t} \; & \sum_{i=1}^{m} \boldsymbol{\alpha}_i^\top \mathbf{e}_{y_i} + \Lambda t \\ \text{s.t. } & \forall k,\ t \leq -\frac{C}{2} \sum_{i,j=1}^{m} \boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_j K_k(x_i, x_j) \\ & \qquad - \sum_{i=1}^{m} \sum_{y \neq y_i} \boldsymbol{\beta}_{i,y} \eta_k(x_i, y_i, y) \\ & \forall i,\ \sum_{y \neq y_i} \beta_{i,y} = \Gamma,\ \boldsymbol{\beta} \geq \mathbf{0} \\ & \forall i,\ \boldsymbol{\alpha}_i \leq \mathbf{e}_{y_i},\ \boldsymbol{\alpha}_i^\top \mathbf{1} = 0\,.\end{aligned}$$

Table 4. Performance of several algorithms on the *caltech101* dataset for varying numbers of training points per class (PPC). The dataset consists of 102 classes and 48 kernels.

PPC	UNIF	BINMKL	OBSC	UFO	M^3K
5	46.0 ± 0.9	54.0 ± 0.7	52.5 ± 0.6	47.9 ± 0.7	51.4 ± 1.2
10	57.9 ± 0.8	65.9 ± 0.9	65.1 ± 1.0	62.9 ± 0.7	66.0 ± 1.1
15	64.6 ± 1.0	71.8 ± 0.5	71.4 ± 0.6	70.6 ± 0.6	72.6 ± 0.9
20	68.4 ± 0.9	75.4 ± 1.1	75.7 ± 1.2	73.5 ± 1.1	76.0 ± 0.8
25	71.9 ± 1.6	77.1 ± 1.2	78.2 ± 0.8	75.5 ± 1.3	79.3 ± 1.0

G. Caltech 101 performance

Here we present the numerical values that are displayed in figure 2.