# A PROOFS

## A.1 PROOF OF THEOREM 4.3

The theorem is due to (Pelckmans and Suykens, 2007), but no full proof could be found in the literature so we supply one here. The proof follows the familiar strategy of using a McDiarmid-type inequality followed by the introduction of a ghost sample, requiring a little more manipulation due to the transductive setting.

We require some preliminaries: let $\mathcal{P}$ be the set of all $n!$ permutations of $n = m + u$ objects $\mathcal{Z}$: for each $\boldsymbol{\pi} \in \mathcal{P}$, each $\pi_i$ is a distinct element of $\mathcal{Z}$. Let $\boldsymbol{\pi}^{ij}$ be the permutation vector obtained by exchanging element $i$ with $j$ in $\boldsymbol{\pi}$. We use the following lemma.

**Lemma A.1.** *(El-Yaniv and Pechyony, 2007, Lemma 3) Suppose that, for each $\boldsymbol{\pi}$, $f : \mathcal{P} \to \mathbb{R}$ is symmetric on $(\pi_1, ...\pi_m)$ and on $(\pi_{m+1}, ...\pi_n)$ and $|f(\boldsymbol{\pi}) - f(\boldsymbol{\pi}^{ij})| \le \beta$ for all $i$ and $j$. Let $\boldsymbol{\pi}$ be drawn uniformly at random from $\mathcal{P}$, then*

$$\mathbb{P}_{\boldsymbol{\pi}}\left(f(\boldsymbol{\pi}) - \mathbb{E}_{\boldsymbol{\pi}}(f(\boldsymbol{\pi})) \ge \epsilon\right) \le \exp\left(\frac{-2\epsilon^2}{\beta^2 \min(m,u)}\right).$$

We now prove the theorem.

**Proof** Define $D(\mathcal{S}) := \sup_{h\in\mathcal{H}}\left(\text{risk}^{\ell}_{\mathcal{T}}(h) - \widehat{\text{risk}}^{\ell}_{\mathcal{S}}(h)\right)$ and notice that $D$ satisfies the conditions of Lemma A.1 with $\beta = C(\frac{1}{m} + \frac{1}{u})$, thus with probability at least $1 - \delta$ over the draw of $\mathcal{S}$

$$D(\mathcal{S}) \le \mathbb{E}_{\mathcal{S}}(D(\mathcal{S})) + C\left(\frac{1}{m} + \frac{1}{u}\right)\sqrt{\frac{\min(m,u)}{2}\log\frac{1}{\delta}}. \qquad (17)$$

Denote $Z_i := (X_i, Y_i)$ for each $(X_i, Y_i)$ drawn from $\mathcal{Z}$. For each $h \in \mathcal{H}$ denote $\ell_h(Z_i) := \ell(h(X_i), Y_i)$ so that $\mathcal{L}_{\mathcal{H}} := \{\ell_h \ : \ h \in \mathcal{H}\}$ is the class of loss functions indexed by $\mathcal{H}$ over $\mathcal{Z}$. We have

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}(D(\mathcal{S})) &= \mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left(\text{risk}^{\ell}_{\mathcal{T}}(h) - \widehat{\text{risk}}^{\ell}_{\mathcal{S}}(h)\right)\right] \\
&= \mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{u}\sum_{i=1}^{u}\ell_h(Z_{t_i}) - \frac{1}{m}\sum_{i=1}^{m}\ell_h(Z_{s_i})\right)\right] \qquad (18) \\
&= \mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{u}\sum_{i=1}^{n}\ell_h(\boldsymbol{z}_i) - \left(\frac{1}{m} + \frac{1}{u}\right)\sum_{i=1}^{m}\ell_h(Z_{s_i})\right)\right] \\
&= \frac{n}{u}\mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}\ell_h(\boldsymbol{z}_i) - \frac{1}{m}\sum_{i=1}^{m}\ell_h(Z_{s_i})\right)\right] \\
&= \frac{n}{u}\mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}_{\mathcal{S}'}\left[\frac{1}{m}\sum_{i=1}^{m}\ell_h(Z'_{s_i})\right] \right.\right. \\
&\qquad\qquad \left.\left. - \frac{1}{m}\sum_{i=1}^{m}\ell_h(Z_{s_i})\right)\right] \qquad (19)
\end{aligned}$$

where $\mathcal{S}' = \{Z'_{s_1}, ...Z'_{s_m}\} = \{(X'_{s_1}, Y'_{s_1}), ...(X'_{s_m}, Y'_{s_m})\}$ is a familiar "ghost sample" drawn according to the same distribution as $\mathcal{S}$, that is, uniformly without replacement from $\mathcal{Z}$. Continuing, the r.h.s. of (19) is no greater than,

$$\begin{aligned}
&\frac{n}{u}\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\ell_h(Z'_{s_i}) - \ell_h(Z_{s_i})\right)\right] \\
&\le \frac{n}{u}\mathbb{E}_{\mathcal{S},\mathcal{S}',\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\ell_h(\widetilde{Z}'_{s_i}) - \ell_h(\widetilde{Z}_{s_i})\right)\right] \qquad (20) \\
&= \frac{n}{u}\mathbb{E}_{\mathcal{S},\mathcal{S}',\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\ell_h(Z'_{s_i}) - \sigma_i\ell_h(Z_{s_i})\right)\right], \\
&\qquad\qquad (21)
\end{aligned}$$

where the $\{\sigma_i\}_{i=1}^{m}$ are independent Rademacher variables and where $\widetilde{Z}_{s_i} := \frac{1}{2}(1+\sigma_i)Z_{s_i} + \frac{1}{2}(1-\sigma_i)Z'_{s_i}$ and $\widetilde{Z}'_{s_i} := \frac{1}{2}(1-\sigma_i)Z_{s_i} + \frac{1}{2}(1+\sigma_i)Z'_{s_i}$.

Inequality in (20) occurs because $\widetilde{\mathcal{S}} := \{\widetilde{Z}_{s_1}, ...\widetilde{Z}_{s_m}\}$ and $\widetilde{\mathcal{S}}' := \{\widetilde{Z}'_{s_1}, ...\widetilde{Z}'_{s_m}\}$ can each contain repeated instances and are less likely than $\mathcal{S}$ and $\mathcal{S}'$ to have in common a copy of the same labeled point, thus the expected supremum is larger: we prove this formally, for a particular $\boldsymbol{\sigma}$, $\mathcal{S}$ and $\mathcal{S}'$ denote,

$$\begin{aligned}
\mathcal{K} &:= \{(i,j) \ : \ Z_{s_i} = Z'_{s_j}\} \\
\widetilde{\mathcal{K}} &:= \{(i,j) \ : \ \widetilde{Z}_{s_i} = \widetilde{Z}'_{s_j}\},
\end{aligned}$$

and call such occurances "clashes". Put $N := |\mathcal{K}| - |\widetilde{\mathcal{K}}| \ge 0$ so that the action of $\boldsymbol{\sigma}$ on $\mathcal{S}$, $\mathcal{S}'$ swaps $N$ clashes; there are $N$ instances which $\mathcal{S}$ and $\mathcal{S}'$ had in common, which occur in one of $\widetilde{\mathcal{S}}$, $\widetilde{\mathcal{S}}'$ with multiplicity 2. Now let $m_0 = m - |\mathcal{K}|$ and define,

$$\begin{aligned}
\Sigma &:= \{\zeta_1, ...\zeta_{m_0}\} := \mathcal{S}\backslash\{Z_i : (i,j) \in \mathcal{K} \text{ for some } j\} \\
\Sigma' &:= \{\zeta'_1, ...\zeta'_{m_0}\} := \mathcal{S}'\backslash\{Z'_j : (i,j) \in \mathcal{K} \text{ for some } i\} \\
\widetilde{\Sigma} &:= \{\widetilde{\zeta}_1, ...\widetilde{\zeta}_{m_0+N}\} := \widetilde{\mathcal{S}}\backslash\{\widetilde{Z}_i : (i,j) \in \widetilde{\mathcal{K}} \text{ for some } j\} \\
\widetilde{\Sigma}' &:= \{\widetilde{\zeta}'_1, ...\widetilde{\zeta}'_{m_0+N}\} := \widetilde{\mathcal{S}}'\backslash\{\widetilde{Z}'_j : (i,j) \in \widetilde{\mathcal{K}} \text{ for some } i\}
\end{aligned}$$

so that, for example, $\Sigma$ is $\mathcal{S}$ with any elements common to $\mathcal{S}$ and $\mathcal{S}'$ removed. Note that $\{\zeta_1, ...\zeta_{m_0}, \zeta'_1, ...\zeta'_{m_0}\}$ are all distinct. Further, w.l.o.g. we order $\widetilde{\Sigma}$ and $\widetilde{\Sigma}'$ such that at least one copy of any elements which occur in either $\widetilde{\Sigma}$ or $\widetilde{\Sigma}'$ with multiplicity 2 (there are $N$ such elements in total, shared between $\widetilde{\Sigma}$ and $\widetilde{\Sigma}'$) is placed in a position $j$ where $m_0 < j \le m_0 + N$. This ordering ensures $\{\widetilde{\zeta}_1, ...\widetilde{\zeta}_{m_0}, \widetilde{\zeta}'_1, ...\widetilde{\zeta}'_{m_0}\}$ are all distinct. Because of this, the sets $\{\zeta_1, ...\zeta_{m_0}, \zeta'_1, ...\zeta'_{m_0}\}$ and $\{\widetilde{\zeta}_1, ...\widetilde{\zeta}_{m_0}, \widetilde{\zeta}'_1, ...\widetilde{\zeta}'_{m_0}\}$ have the same distribution: they are both drawn uniformly without replacement from $\mathcal{Z}$. Now we set

$$h^* := \underset{h\in\mathcal{H}}{\text{argmax}}\sum_{i=1}^{m_0}\ell_h(\widetilde{\zeta}'_i) - \ell_h(\widetilde{\zeta}_i)$$

and note,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S},\mathcal{S}',\boldsymbol{\sigma}}\Bigg[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\ell_h(\widetilde{Z}'_{s_i})-\ell_h(\widetilde{Z}_{s_i})\right)\\
&\qquad -\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\ell_h(Z'_{s_i})-\ell_h(Z_{s_i})\right)\Bigg]\\
&=\mathbb{E}_{\mathcal{S},\mathcal{S}',\boldsymbol{\sigma}}\Bigg[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m_0+N}\ell_h(\widetilde{\zeta}'_i)-\ell_h(\widetilde{\zeta}_i)\right)\\
&\qquad -\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m_0}\ell_h(\zeta'_i)-\ell_h(\zeta_i)\right)\Bigg]\\
&\geq\mathbb{E}_{\mathcal{S},\mathcal{S}',\boldsymbol{\sigma}}\Bigg[\frac{1}{m}\sum_{i=1}^{m_0+N}\ell_{h^*}(\widetilde{\zeta}'_i)-\ell_{h^*}(\widetilde{\zeta}_i)\\
&\qquad -\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m_0}\ell_h(\zeta'_i)-\ell_h(\zeta_i)\right)\Bigg]\\
&=\mathbb{E}_{\mathcal{S},\mathcal{S}',\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{i=m_0+1}^{m_0+N}\ell_{h^*}(\widetilde{\zeta}'_i)-\ell_{h^*}(\widetilde{\zeta}_i)\right]\\
&\geq 0.
\end{aligned}$$

The final line holds because, conditional on $\{\widetilde{\zeta}_1,...\widetilde{\zeta}_{m_0},\widetilde{\zeta}'_1,...\widetilde{\zeta}'_{m_0}\}$, elements of $\{\widetilde{\zeta}'_{m_0+1},...\widetilde{\zeta}'_{m_0+N}\}$ are drawn from $\mathcal{Z}\backslash\{\widetilde{\zeta}_1,...\widetilde{\zeta}_{m_0}\}$ and elements of $\{\widetilde{\zeta}_{m_0+1},...\widetilde{\zeta}_{m_0+N}\}$ are drawn from $\mathcal{Z}\backslash\{\widetilde{\zeta}_1,...\widetilde{\zeta}_{m_0}\}$. Thus (20) holds.

To continue, we finally just note,

$$\begin{aligned}
(21) &\leq 2\frac{n}{u}\mathcal{R}^{\mathrm{trs}}_m(\mathcal{L}_{\mathcal{H}})\\
&\leq 2K\frac{n}{u}\mathcal{R}^{\mathrm{trs}}_m(\mathcal{H}),
\end{aligned}$$

The final line is a consequence of the contraction inequality for Rademacher complexities, (Meir and Zhang, 2003, Theorem 7).

Finally, notice the symmetry in (18) for $m \leftrightarrow u$ and that by producing precisely the symmetrically opposite argument we would derive $\mathbb{E}_{\mathcal{S}}(D(\mathcal{S})) \leq 2K\frac{n}{m}\mathcal{R}^{\mathrm{trs}}_u(\mathcal{H})$, hence $\mathbb{E}_{\mathcal{S}}(D(\mathcal{S})) \leq \frac{2Kn}{\max(m,u)}\mathcal{R}^{\mathrm{trs}}_{\min(m,u)}(\mathcal{H})$. □

### A.2 PROOF OF THEOREM 5.1

**Proof** Let $\mathcal{T} := \{(\boldsymbol{X}_{t_1},Y_{t_1}),...(\boldsymbol{X}_{t_u},Y_{t_u})\}$ where the $Y_{t_i}$ are drawn from the conditional $P_{Y|X}$. The transductive bound Theorem 4.3 implies that,

$$\mathbb{P}\left(\sup_{h\in\widetilde{\mathcal{H}}_\beta}\left(\mathrm{risk}^{\ell}_{\mathcal{S}\cup\mathcal{T}}(h)-\widehat{\mathrm{risk}}^{\ell}_{\mathcal{S}}(h)\right)\leq 2K\mathcal{R}^{\mathrm{trs}}_m(\widetilde{\mathcal{H}}_\beta)+C\sqrt{\frac{1}{2m}\log\frac{2}{\delta}}\right)\geq 1-\frac{\delta}{2},$$

the empirical counterpart of Theorem 2.1 (see e.g. Boucheron et al., 2005) implies that

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}_\alpha}\left(\mathrm{risk}^{\ell}(h)-\mathrm{risk}^{\ell}_{\mathcal{S}\cup\mathcal{T}}(h)\right)\leq 2K\widehat{\mathcal{R}}^{\mathrm{ind}}_{\mathcal{I}}(\mathcal{H}_\alpha)+3C\sqrt{\frac{1}{2n}\log\frac{4}{\delta}}\right)\geq 1-\frac{\delta}{2},$$

and (16) follows from the union bound. The final results follow from the bounds (8,2) on the transductive Rademacher complexity and empirical inductive Rademacher complexity. □

## B STRUCTURE-DEPENDENT RISK BOUND AND REGULARIZATION

Theorem 3.2 supplies a risk bound in terms of the observed cluster structure in the training sample.

**Theorem B.1.** *Using the notation of Theorem 3.2, and when $\ell(\cdot,\cdot)$ is positive and bounded by $C$, for all $h\in\mathcal{H}$,*

$$\mathbb{P}_{\mathcal{S}}\left(\mathrm{risk}^{\ell}(h)\leq\widehat{\mathrm{risk}}^{\ell}_{\mathcal{S}}(h)+2K\left(B\sqrt{\frac{|\mathcal{C}|}{m}}+2\sqrt{\frac{2F'(h)\rho_{\mathcal{S}}}{m\kappa}}\right)+3C\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right)\geq 1-\delta.$$

*where $F'(h) := \min_{r\in\{1,2,...\}}\max\left(\alpha_r,\frac{r+1}{r}F(h)\right)$ and $\alpha_r := \frac{9C^2\kappa r\log 2}{16K^2\rho_{\mathcal{S}}}$.*

**Proof** Define the stratification: $\mathcal{H}^{(0)} = \{\}$ and, for $t\in\{1,2,...\}$, $\mathcal{H}^{(t)} = \mathcal{H}_{\alpha_t}$. The empirical version of Theorem 2.1 (see e.g. (Boucheron et al., 2005)) implies that with probability at least $1-\frac{\delta}{2^t}$ simultaneously for all $h\in\mathcal{H}^{(t)}\backslash\mathcal{H}^{(t-1)}$ we have,

$$\begin{aligned}
&\mathrm{risk}^{\ell}(h)-\widehat{\mathrm{risk}}^{\ell}_{\mathcal{S}}(h)\leq 2K\widehat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H}_{\alpha_t})+3C\sqrt{\frac{\log\frac{2^{t+1}}{\delta}}{2m}}\\
&\leq 2K\left(B\sqrt{\frac{|\mathcal{C}|}{m}}+\sqrt{\frac{2\alpha_t\rho_{\mathcal{S}}}{m\kappa}}\right)\\
&\qquad +3C\sqrt{\frac{t\log 2}{2m}}+3C\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\\
&\leq 2K\left(B\sqrt{\frac{|\mathcal{C}|}{m}}+2\sqrt{\frac{2\alpha_t\rho_{\mathcal{S}}}{m\kappa}}\right)+3C\sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \qquad (22)
\end{aligned}$$

Now noting that for $r\in\{1,2,...\}$, $\alpha_t>\alpha_r$ implies that $t\geq r+1$ and $\alpha_t\leq\frac{r+1}{r}\alpha_{t-1}$ so

$$\alpha_t\leq\min_{r\in\{1,2,...\}}\max\left(\alpha_r,\frac{r+1}{r}\alpha_{t-1}\right)\leq F'(h) \qquad (23)$$

The result then follows by combining (23) with (22) and applying the union bound over all $t \in \{1, 2, ...\}$.
$\square$

Theorem B.1 suggests an algorithm: pick the classifier which minimizes the bound. this is simply regularization w.r.t. the complexity $F(\cdot)$ but the regularization parameters are determined by the observed cluster structure in the data. In principle the information needed to choose the regularization parameter should be encoded in the data, so it would be of interest to understand this relationship and reduce the need for cross validation.

A special case of the above is RKHS regularization, obtained by picking the 1-strongly convex Hilbert space norm as a complexity, $F(h) = \frac{1}{2}||h||^2_{\mathcal{H}}$. The cluster structure in this case is that in feature space.