Convex Calibrated Surrogates for Hierarchical Classification
Supplementary Material

A. Additional Notation and Setup

Let μ be the marginal distribution induced by D over $\mathcal{X}$, and let $\mathbf{p}(x)$ be the distribution over $[n]$ conditioned on $X = x$. For every function $\ell : [n] \times [k] \to \mathbb{R}_+$ and $t \in [k]$ let $\boldsymbol{\ell}_t = [\ell(1,t), \ldots, \ell(n,t)]^\top \in \mathbb{R}^n_+$. For every surrogate $\psi : [n] \times \mathbb{R}^d \to \mathbb{R}_+$ let $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^n_+$ be a vector function such that $\psi_y(\mathbf{u}) = \psi(y, \mathbf{u})$ for $y \in [n], \mathbf{u} \in \mathbb{R}^d$. For any integer $d' \in \mathbb{Z}_+$ and pair of vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{d'}$, their inner product is denoted as $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^{d'} u_i v_i$. For a vector $\mathbf{u} \in \mathbb{R}^n$ and a positive integer $a \le n$, the vector $\mathbf{u}\big|_{1:a} \in \mathbb{R}^a$ gives the first a components of $\mathbf{u}$.

Define the conditional regrets $\mathbf{R}^{\ell^H}_{\mathbf{p}}$, $\mathbf{R}^{\ell^{?,n}}_{\mathbf{p}}$ and $\mathbf{R}^{\psi}_{\mathbf{p}}$ as the regrets incurred for a singleton instance space $\mathcal{X}$, with conditional probability $\mathbf{p} \in \Delta_n$. In particular, we have that

$$\begin{aligned}
\mathbf{R}^{\ell^H}_{\mathbf{p}}[\widehat{y}] &= \langle \mathbf{p}, \boldsymbol{\ell}^H_{\widehat{y}} \rangle - \inf_{y' \in [n]} \langle \mathbf{p}, \boldsymbol{\ell}^H_{y'} \rangle, && \forall \widehat{y} \in [n] \\
\mathbf{R}^{\ell^{?,n}}_{\mathbf{p}}[\widehat{y}] &= \langle \mathbf{p}, \boldsymbol{\ell}^{?,n}_{\widehat{y}} \rangle - \inf_{y' \in [n] \cup \{\perp\}} \langle \mathbf{p}, \boldsymbol{\ell}^{?,n}_{y'} \rangle, && \forall \widehat{y} \in [n] \cup \{\perp\} \\
\mathbf{R}^{\psi}_{\mathbf{p}}[\mathbf{u}] &= \langle \mathbf{p}, \boldsymbol{\psi}(\mathbf{u}) \rangle - \inf_{\mathbf{u}' \in \mathbb{R}^d} \langle \mathbf{p}, \boldsymbol{\psi}(\mathbf{u}') \rangle, && \forall \mathbf{u} \in \mathbb{R}^d .
\end{aligned}$$

Let μ be the marginal distribution induced by D over $\mathcal{X}$, and let $\mathbf{p}(x)$ be the distribution over $[n]$ conditioned on $X = x$. Then we have by linearity of expectation that,

$$\begin{aligned}
\mathbf{R}^{\ell^H}_{D}[g] &= \mathbf{E}_{X \sim \mu} \mathbf{R}^{\ell^H}_{\mathbf{p}(X)}[g(X)] && (4) \\
\mathbf{R}^{\ell^{?,n}}_{D}[g'] &= \mathbf{E}_{X \sim \mu} \mathbf{R}^{\ell^{?,n}}_{\mathbf{p}(X)}[g'(X)] && (5) \\
\mathbf{R}^{\psi}_{D}[\mathbf{f}] &= \mathbf{E}_{X \sim \mu} \mathbf{R}^{\psi}_{\mathbf{p}(X)}[\mathbf{f}(X)] . && (6)
\end{aligned}$$

For all $0 \le j \le h$, define $\ell^{H,j} : [n_j] \times [n_j] \to \mathbb{R}_+$ as simply the restriction of ℓ^H to $[n_j] \times [n_j]$.

B. Proofs

B.1. Proof of Theorem 1

Theorem. *Let $H = ([n], E, W)$ and let $\ell^H : [n] \times [n] \to \mathbb{R}_+$ be the tree-distance loss for the tree H. Let $\mathbf{p} \in \Delta_n$, and $y \in [n]$. Then there exists a $g^* : \mathcal{X} \to [n]$ such that for all $x \in \mathcal{X}$ the following holds:*

(a) $S_{g^*(x)}(\mathbf{p}(x)) \ge \frac{1}{2}$

(b) $S_y(\mathbf{p}(x)) \le \frac{1}{2}, \forall y \in C(g^*(x))$.

Also, g^* is a Bayes optimal classifier for the tree distance loss, i.e.

$$\mathbf{R}^{\ell^H}_{D}[g^*] = 0 .$$

Proof. We shall simply show for all $\mathbf{p} \in \Delta_n$, there exists a $y^* \in [n]$ such that

$$\begin{aligned}
S_{y^*}(\mathbf{p}) &\ge \frac{1}{2} && (7) \\
S_y(\mathbf{p}) &\le \frac{1}{2}, \qquad \forall y \in C(y^*) , && (8)
\end{aligned}$$

and is such that

$$\langle \mathbf{p}, \boldsymbol{\ell}^H_{y^*} \rangle = \min_{y \in [n]} \langle \mathbf{p}, \boldsymbol{\ell}^H_y \rangle \ .$$

This would imply $\mathbf{R}^{\ell^H}_{\mathbf{p}}[y^*] = 0$. The theorem then simply follows from linearity of expectation using Equation 4.

Let $\mathbf{p} \in \Delta_n$. We construct a $y^* \in [n]$ satisfying Equations 7 and 8 in the following way. We start at the root node, which always satisfies Equation 7, and keep on moving to the child of the current node that satisfies Equation 7, and terminate when we reach a leaf node, or a node where all of its children fail Equation 7. Clearly the resulting node, y^*, satisfies both Equations 7 and 8.

Now we show that y^* indeed minimizes $\langle \mathbf{p}, \boldsymbol{\ell}^H_y \rangle$ over $y \in [n]$.

Let $y' \in \text{argmin}_t \langle \mathbf{p}, \boldsymbol{\ell}^H_t \rangle$. If $y' = y^*$ we are done, hence assume $y' \neq y^*$.

Case 1: $y' \notin D(y^*)$

$$\begin{aligned}
\langle \mathbf{p}, \boldsymbol{\ell}^H_{y'} \rangle - \langle \mathbf{p}, \boldsymbol{\ell}^H_{y^*} \rangle &= \sum_{y \in D(y^*)} p_y(\ell^H(y, y') - \ell^H(y, y^*)) + \sum_{y \in [n] \setminus D(y^*)} p_y(\ell^H(y, y') - \ell^H(y, y^*)) \\
&= \sum_{y \in D(y^*)} p_y(\ell^H(y^*, y')) + \sum_{y \in [n] \setminus D(y^*)} p_y(\ell^H(y, y') - \ell^H(y, y^*)) \\
&\geq \sum_{y \in D(y^*)} p_y(\ell^H(y^*, y')) + \sum_{y \in [n] \setminus D(y^*)} p_y(-\ell^H(y', y^*))) \\
&= \ell^H(y', y^*)(2S_{y^*}(\mathbf{p}) - 1) \\
&\geq 0
\end{aligned}$$

Case 2: $y' \in D(y^*) \setminus C(y^*)$

Let $\hat{y}$ be the child of y^* that is the ancestor of y'. Hence we have $S_{\hat{y}}(\mathbf{p}) \leq \frac{1}{2}$.

$$\begin{aligned}
\langle \mathbf{p}, \boldsymbol{\ell}^H_{y'} \rangle - \langle \mathbf{p}, \boldsymbol{\ell}^H_{y^*} \rangle &= \sum_{y \in D(\hat{y})} p_y(\ell^H(y, y') - \ell^H(y, y^*)) + \sum_{y \in [n] \setminus D(\hat{y})} p_y(\ell^H(y, y') - \ell^H(y, y^*)) \\
&= \sum_{y \in D(\hat{y})} p_y(\ell^H(y, y') - \ell^H(y, y^*)) + \sum_{y \in [n] \setminus D(\hat{y})} p_y(\ell^H(y^*, y')) \\
&\geq \sum_{y \in D(\hat{y})} p_y(-\ell^H(y^*, y')) + \sum_{y \in [n] \setminus D(\hat{y})} p_y(\ell^H(y^*, y')) \\
&= \ell^H(y', y^*)(1 - 2S_{\hat{y}}(\mathbf{p})) \\
&\geq 0
\end{aligned}$$

Case 3: $y' \in C(y^*)$

$$\begin{aligned}
\langle \mathbf{p}, \boldsymbol{\ell}^H_{y'} \rangle - \langle \mathbf{p}, \boldsymbol{\ell}^H_{y^*} \rangle &= \sum_{y \in D(y')} p_y(\ell^H(y, y') - \ell^H(y, y^*)) + \sum_{y \in [n] \setminus D(y')} p_y(\ell^H(y, y') - \ell^H(y, y^*)) \\
&= \sum_{y \in D(y')} p_y(-\ell^H(y', y^*)) + \sum_{y \in [n] \setminus D(y')} p_y(\ell^H(y', y^*)) \\
&= \ell^H(y', y^*)(1 - 2S_{y'}(\mathbf{p})) \\
&\geq 0
\end{aligned}$$

Putting all three cases together we have

$$\langle \mathbf{p}, \boldsymbol{\ell}^H_{y^*} \rangle \leq \langle \mathbf{p}, \boldsymbol{\ell}^H_{y'} \rangle = \min_{y \in [n]} \langle \mathbf{p}, \boldsymbol{\ell}^H_y \rangle \ .$$

□

B.2. Proof of Lemma 3

We first give the proof of a stronger version of Lemma 3.

Lemma 8. *For all* $\mathbf{p} \in \Delta_n, \mathbf{u} \in \mathbb{R}^d$ *we have*

$$\mathbf{R}^{\ell^H}_{\mathbf{p}}[\Upsilon^{\text{cas}}(\mathbf{u})] \leq \sum_{j=1}^{h} \gamma_j(\mathbf{u}_j) \cdot \mathbf{R}^{\ell^{?,n_j}}_{A^j(\mathbf{p})}[\Upsilon^?_j(\mathbf{u}_j)],$$

where $\gamma_j(\mathbf{u}_j) = \begin{cases} 2\alpha_j & \text{if } \Upsilon^?_j(\mathbf{u}_j) \neq \perp \\ 2\beta_j & \text{if } \Upsilon^?_j(\mathbf{u}_j) = \perp \end{cases}.$

Proof. For all $j \in [h]$, we will first prove a bound relating the tree-distance regret at level j, with the tree-distance regret at level $j-1$ and abstain loss regret at level j as follows:

$$\mathbf{R}^{\ell^{H,j}}_{A^j(\mathbf{p})}[\Upsilon^{\text{cas}}_j(\mathbf{u})] \quad \leq \quad \mathbf{R}^{\ell^{H,j-1}}_{A^{j-1}(\mathbf{p})}[\Upsilon^{\text{cas}}_{j-1}(\mathbf{u})] + \gamma_j(\mathbf{u}_j) \cdot \mathbf{R}^{\ell^{?,n_j}}_{A^j(\mathbf{p})}[\Upsilon^?_j(\mathbf{u}_j)] \ .$$

The theorem would simply follow from applying such a bound recursively and observing that $\mathbf{R}^{\ell^{H,0}}_{A^0(\mathbf{p})}[\Upsilon^{\text{cas}}_0(\mathbf{u})] = 0$.

One observation of the tree-distance loss that will be often of use in the proof is the following:

$$\ell^H(y, y') - \ell^H(P(y), y') = \begin{cases} -\ell^H(y, P(y)) & \text{if } y' \in D(y) \\ \ell^H(y, P(y)) & \text{otherwise} \end{cases}$$

The details of the proof follows: Fix $j \in [h], \mathbf{u} \in \mathbb{R}^d, \mathbf{p} \in \Delta_n$.

Let $y^*_j = \text{argmin}_{y \in [n_j]} \mathbf{R}^{\ell^{H,j}}_{A^j(\mathbf{p})}[y]$.

Case 1: $\Upsilon^?_j(\mathbf{u}_j) \neq \perp$

$$\begin{aligned} \mathbf{R}^{\ell^{H,j}}_{A^j(\mathbf{p})}[\Upsilon^{\text{cas}}_j(\mathbf{u})] &= \sum_{y=1}^{n_j} A^j_y(\mathbf{p})(\ell^H(y, \Upsilon^{\text{cas}}_j(\mathbf{u})) - \ell^H(y, y^*_j)) \\ &\leq \ell^H(y^*_j, \Upsilon^{\text{cas}}_j(\mathbf{u}))(1 - 2A^j_{\Upsilon^{\text{cas}}_j(\mathbf{u})}(\mathbf{p})) \end{aligned} \tag{9}$$

We also have,

$$\begin{aligned} \mathbf{R}^{\ell^{?,n_j}}_{A^j(\mathbf{p})}[\Upsilon^?_j(\mathbf{u}_j)] &= 1 - A^j_{\Upsilon^?_j(\mathbf{u}_j)}(\mathbf{p}) - \min_{y \in [n_j] \cup \{\perp\}} \langle A^j(\mathbf{p}), \boldsymbol{\ell}^{?,n_j}_y \rangle \\ &\geq 1 - A^j_{\Upsilon^?_j(\mathbf{u}_j)}(\mathbf{p}) - \langle A^j(\mathbf{p}), \boldsymbol{\ell}^{?,n_j}_\perp \rangle \\ &= \frac{1}{2} - A^j_{\Upsilon^?_j(\mathbf{u}_j)}(\mathbf{p}) \\ &= \frac{1}{2} - A^j_{\Upsilon^{\text{cas}}_j(\mathbf{u})}(\mathbf{p}) \ . \end{aligned} \tag{10}$$

The last inequality above follows because if $\Upsilon^?_j(\mathbf{u}_j) \neq \perp$, then $\Upsilon^{\text{cas}}_j(\mathbf{u}) = \Upsilon^?_j(\mathbf{u}_j)$.

Putting Equations 9 and 10 together, we get

$$\begin{aligned}
\mathbf{R}^{\ell^{H,j}}_{A^j(\mathbf{p})}[\Upsilon^{\text{cas}}_j(\mathbf{u})] &\leq 2\ell^H(y^*_j, \Upsilon^{\text{cas}}_j(\mathbf{u})) \cdot \mathbf{R}^{\ell^{?,n_j}}_{A^j(\mathbf{p})}[\Upsilon^?_j(\mathbf{u}_j)] \\
&\leq 2\alpha_j \cdot \mathbf{R}^{\ell^{?,n_j}}_{A^j(\mathbf{p})}[\Upsilon^?_j(\mathbf{u}_j)] \qquad (11)
\end{aligned}$$

Case 2: $\Upsilon^?_j(\mathbf{u}_j) = \perp$

In this case $\Upsilon^{\text{cas}}_j(\mathbf{u}) = \Upsilon^{\text{cas}}_{j-1}(\mathbf{u})$, and hence $\text{lev}(\Upsilon^{\text{cas}}_j(\mathbf{u})) \leq j - 1$.

We now have,

$$\begin{aligned}
\langle A^j(\mathbf{p}), \boldsymbol{\ell}^{H,j}_{\Upsilon^{\text{cas}}_j(\mathbf{u})}\rangle - \langle A^{j-1}(\mathbf{p}), \boldsymbol{\ell}^{H,j-1}_{\Upsilon^{\text{cas}}_{j-1}(\mathbf{u})}\rangle &= \langle A^j(\mathbf{p}), \boldsymbol{\ell}^{H,j}_{\Upsilon^{\text{cas}}_j(\mathbf{u})}\rangle - \langle A^{j-1}(\mathbf{p}), \boldsymbol{\ell}^{H,j-1}_{\Upsilon^{\text{cas}}_j(\mathbf{u})}\rangle \\
&= \sum_{y \in N_{=j}} S_y(\mathbf{p}) \left(\ell^H(y, \Upsilon^{\text{cas}}_j(\mathbf{u})) - \ell^H(P(y), \Upsilon^{\text{cas}}_j(\mathbf{u}))\right) \\
&= \sum_{y \in N_{=j}} S_y(\mathbf{p}) \ell^H(y, P(y)) \qquad (12)
\end{aligned}$$

For ease of analysis, we divide case 2, further into two sub-cases.

Case 2a: $\text{lev}(y^*_j) < j$

$$\begin{aligned}
\langle A^{j-1}(\mathbf{p}), \boldsymbol{\ell}^{H,j-1}_{y^*_{j-1}}\rangle - \langle A^j(\mathbf{p}), \boldsymbol{\ell}^{H,j}_{y^*_j}\rangle &= \langle A^{j-1}(\mathbf{p}), \boldsymbol{\ell}^{H,j-1}_{y^*_{j-1}}\rangle - \langle A^{j-1}(\mathbf{p}), \boldsymbol{\ell}^{H,j-1}_{y^*_j}\rangle \\
&\quad + \langle A^{j-1}(\mathbf{p}), \boldsymbol{\ell}^{H,j-1}_{y^*_j}\rangle - \langle A^j(\mathbf{p}), \boldsymbol{\ell}^{H,j}_{y^*_j}\rangle \\
&\leq \langle A^{j-1}(\mathbf{p}), \boldsymbol{\ell}^{H,j-1}_{y^*_j}\rangle - \langle A^j(\mathbf{p}), \boldsymbol{\ell}^{H,j}_{y^*_j}\rangle \\
&= \sum_{y \in N_{=j}} S_y(\mathbf{p})(\ell^H(P(y), y^*_j) - \ell^H(y, y^*_j)) \\
&= \sum_{y \in N_{=j}} S_y(\mathbf{p})(-\ell^H(y, P(y))) \qquad (13)
\end{aligned}$$

Adding, Equation 12 and 13, we get

$$\mathbf{R}^{\ell^{H,j}}_{A^j(\mathbf{p})}[\Upsilon^{\text{cas}}_j(\mathbf{u})] \leq \mathbf{R}^{\ell^{H,j-1}}_{A^{j-1}(\mathbf{p})}[\Upsilon^{\text{cas}}_{j-1}(\mathbf{u})] \qquad (14)$$

Case 2b: $\text{lev}(y^*_j) = j$

$$\begin{aligned}
\langle A^{j-1}(\mathbf{p}), \boldsymbol{\ell}^{H,j-1}_{y^*_{j-1}}\rangle - \langle A^{j-1}(\mathbf{p}), \boldsymbol{\ell}^{H}_{y^*_j}\big|_{[1:n_{j-1}]}\rangle &\leq \langle A^{j-1}(\mathbf{p}), \boldsymbol{\ell}^{H,j-1}_{P(y^*_j)}\rangle - \langle A^{j-1}(\mathbf{p}), \boldsymbol{\ell}^{H}_{y^*_j}\big|_{[1:n_{j-1}]}\rangle \\
&= \sum_{y \in N_{\leq j-1}} A^{j-1}_y(\mathbf{p})(\ell^H(y, P(y^*_j)) - \ell^H(y, y^*_j)) \\
&= \sum_{y \in N_{\leq j-1}} A^{j-1}_y(\mathbf{p})(-\ell^H(y^*_j, P(y^*_j))) \\
&= -\ell^H(y^*_j, P(y^*_j)) \qquad (15)
\end{aligned}$$

Also,

$$\begin{aligned}
\langle A^{j-1}(\mathbf{p}), \boldsymbol{\ell}^{H}_{y^*_j}\big|_{[1:n_{j-1}]}\rangle - \langle A^j(\mathbf{p}), \boldsymbol{\ell}^{H,j}_{y^*_j}\rangle &= \sum_{y \in N_{=j}} S_y(\mathbf{p})(\ell^H(P(y), y^*_j) - \ell^H(y, y^*_j)) \\
&= \sum_{y \in N_{=j} \setminus \{y^*_j\}} S_y(\mathbf{p})(-\ell^H(y, P(y))) + S_{y^*_j}(\mathbf{p})(\ell^H(y^*_j, P(y^*_j))) \,. \qquad (16)
\end{aligned}$$

Adding Equations 12, 15 and 16, we get

$$\begin{aligned}\mathbf{R}^{\ell^{H,j}}_{A^j(\mathbf{p})}[\Upsilon^{\text{cas}}_j(\mathbf{u})] &\leq \mathbf{R}^{\ell^{H,j-1}}_{A^{j-1}(\mathbf{p})}[\Upsilon^{\text{cas}}_{j-1}(\mathbf{u})] + (2S_{y^*_j}(\mathbf{p})-1)\cdot \ell^H(y^*_j, P(y^*_j)) \\ &\leq \mathbf{R}^{\ell^{H,j-1}}_{A^{j-1}(\mathbf{p})}[\Upsilon^{\text{cas}}_{j-1}(\mathbf{u})] + (2S_{y^*_j}(\mathbf{p})-1)\cdot \beta_j \ . \qquad (17)\end{aligned}$$

Inequality 17 follows because by the definitions of y^*_j and Theorem 1 , we have $S_{y^*_j}(\mathbf{p}) \geq \frac{1}{2}$.

Also, we have that

$$\begin{aligned}\mathbf{R}^{\ell^{?,n_j}}_{A^j(\mathbf{p})}[\Upsilon^?_j(\mathbf{u}_j)] &= \mathbf{R}^{\ell^{?,n_j}}_{A^j(\mathbf{p})}[\perp] \\ &= \frac{1}{2} - \min_{y\in[n]\cup\{\perp\}} \langle A^j(\mathbf{p}), \boldsymbol{\ell}^{?,n_j}_y\rangle \\ &\geq \frac{1}{2} - \langle A^j(\mathbf{p}), \boldsymbol{\ell}^{?,n_j}_{y^*_j}\rangle \\ &= \frac{1}{2} - (1 - S_{y^*_j}(\mathbf{p})) \\ &= S_{y^*_j}(\mathbf{p}) - \frac{1}{2} \ . \qquad (18)\end{aligned}$$

Putting Equations 17 and 18 together, we have that

$$\mathbf{R}^{\ell^{H,j}}_{A^j(\mathbf{p})}[\Upsilon^{\text{cas}}_j(\mathbf{u})] \leq \mathbf{R}^{\ell^{H,j-1}}_{A^{j-1}(\mathbf{p})}[\Upsilon^{\text{cas}}_{j-1}(\mathbf{u})] + 2\beta_j \cdot \mathbf{R}^{\ell^{?,n_j}}_{A^j(\mathbf{p})}[\Upsilon^?_j(\mathbf{u}_j)]. \qquad (19)$$

Putting the results for case 1, case 2a and case 2b, from Equations 11, 14 and 19 respectively, we have

$$\mathbf{R}^{\ell^{H,j}}_{A^j(\mathbf{p})}[\Upsilon^{\text{cas}}_j(\mathbf{u})] \leq \mathbf{R}^{\ell^{H,j-1}}_{A^{j-1}(\mathbf{p})}[\Upsilon^{\text{cas}}_{j-1}(\mathbf{u})] + \gamma_j(\mathbf{u}_j) \cdot \mathbf{R}^{\ell^{?,n_j}}_{A^j(\mathbf{p})}[\Upsilon^?_j(\mathbf{u}_j)] \ .$$

□

Now the proof of Lemma 3 follows from certain simple considerations.

Lemma. *For any distribution D over $\mathcal{X} \times [n]$, let $A^j(D)$ be the distribution over $\mathcal{X} \times [n_j]$ given by the distribution of $(X, \text{anc}_j(Y))$ with $(X,Y) \sim D$. For all $j \in [h]$, let $\mathbf{f}_j : \mathcal{X} \to \mathbb{R}^{d_j}$ be such that $\mathbf{f}(x) = [\mathbf{f}_1(x)^\top, \ldots, \mathbf{f}_h(x)^\top]^\top$. Then for all distributions D over $\mathcal{X} \times [n]$ and all functions $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$*

$$\mathbf{R}^{\ell^H}_D[\Upsilon^{\text{cas}} \circ \mathbf{f}] \leq \sum_{j=1}^h 2\alpha_j \cdot \mathbf{R}^{\ell^{?,n_j}}_{A^j(D)}[\Upsilon^?_j \circ \mathbf{f}_j] \ .$$

Proof. Using Lemma 8 and by the observation that $\beta_j \leq \alpha_j$, we have for all $\mathbf{p} \in \Delta_n, \mathbf{u} \in \mathbb{R}^d$ that

$$\mathbf{R}^{\ell^H}_{\mathbf{p}}[\Upsilon^{\text{cas}}(\mathbf{u})] \leq \sum_{j=1}^h 2\alpha_j \cdot \mathbf{R}^{\ell^{?,n_j}}_{A^j(\mathbf{p})}[\Upsilon^?_j(\mathbf{u}_j)] \ .$$

Let $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$ be a function. Then for all $x \in \mathcal{X}$,

$$\mathbf{R}^{\ell^H}_{\mathbf{p}(x)}[\Upsilon^{\text{cas}}(\mathbf{f}(x))] \leq \sum_{j=1}^h 2\alpha_j \cdot \mathbf{R}^{\ell^{?,n_j}}_{A^j(\mathbf{p}(x))}[\Upsilon^?_j(\mathbf{u}_j)] \ .$$

Observe that the the marginal distribution over $\mathcal{X}$ for $A^j(D)$ is exactly the same as for D, while the conditional probability distribution for the distribution $A^j(D)$ at x is exactly equal to $A^j(\mathbf{p}(x))$. The Lemma now immediately follows from linearity of expectation. □

B.3. Proof of Theorem 4

Theorem. *For all $j \in [h]$, let $\psi^j : [n_j] \times \mathbb{R}^{d_j}$ and $\Upsilon^?_j : \mathbb{R}^{d_j} \to n_j$ be such that for all $\mathbf{f}_j : \mathcal{X} \to \mathbb{R}^{d_j}$, and all distributions D over $\mathcal{X} \times [n_j]$ we have*

$$\mathbf{R}_D^{\ell^{?,n_j}}[\Upsilon^?_j \circ \mathbf{f}_j] \leq C \cdot \mathbf{R}_D^{\psi^j}[\mathbf{f}_j],$$

for some constant $C > 0$. Then for all $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$ and distributions D over $\mathcal{X} \times [n]$,

$$\mathbf{R}_D^{\ell^H}[\Upsilon^{\text{cas}} \circ \mathbf{f}] \leq 2\alpha_h C \cdot \mathbf{R}_D^{\psi^{\text{cas}}}[\mathbf{f}] \ .$$

Proof. Fix $\mathbf{u} \in \mathbb{R}^d, \mathbf{p} \in \Delta_n$. From Lemma 3, we have that

$$\begin{aligned}
\mathbf{R}_{\mathbf{p}}^{\ell^H}[\Upsilon^{\text{cas}}(\mathbf{u})] &\leq \sum_{j=1}^{h} 2\alpha_j \cdot \mathbf{R}_{A^j(\mathbf{p})}^{\ell^{?,n_j}}[\Upsilon^?_j(\mathbf{u}_j)] \\
&\leq 2\alpha_h \cdot \sum_{j=1}^{h} \cdot \mathbf{R}_{A^j(\mathbf{p})}^{\ell^{?,n_j}}[\Upsilon^?_j(\mathbf{u}_j)] \\
&\leq 2\alpha_h C \cdot \sum_{j=1}^{h} \mathbf{R}_{A^j(\mathbf{p})}^{\psi^j}[\mathbf{u}_j] \\
&= 2\alpha_h C \cdot \mathbf{R}_{\mathbf{p}}^{\psi^{\text{cas}}}[\mathbf{u}] \ .
\end{aligned}$$

The proof now simply follows from linearity of expectation. □

B.4. Proof of Theorem 6

While Theorem 2 from Ramaswamy et al. (2015) , gives an excess risk bound for the abstain loss excess risk in terms of the OvA-surrogate risk, one can easily get a more refined bound as well from the results of Ramaswamy et al. (2015).

Lemma 9 ((Ramaswamy et al., 2015))**.** *Let $\tau \in (-1, 1)$. For all $\mathbf{u} \in \mathbb{R}^n, \mathbf{p} \in \Delta_n$, and $A = \mathbf{1}(\Upsilon_\tau^{\text{OvA},n}(\mathbf{u}) = n+1)$. Then for all $\mathbf{p} \in \Delta_n$*

$$\mathbf{R}_{\mathbf{p}}^{\ell^{?,n}}[\Upsilon_\tau^{\text{OvA},n}(\mathbf{u})] \quad \leq \quad \left(\frac{\mathbf{1}(\Upsilon^{\text{OvA},n}(\mathbf{u}) = \perp)}{2(1-\tau)} + \frac{\mathbf{1}(\Upsilon^{\text{OvA},n}(\mathbf{u}) \neq \perp)}{2(1+\tau)} \right) \mathbf{R}_{\mathbf{p}}^{\psi^{\text{OvA},n}}[\mathbf{u}] \ .$$

We are now ready to prove Theorem 6.

Theorem. *For $1 \leq j \leq h$, let $\tau_j = \frac{\alpha_j - \beta_j}{\alpha_j + \beta_j}$. Let the component surrogates and predictors of ψ^{cas} and Υ^{cas} be $\psi^j = \psi^{\text{OvA},n_j}$ and $\Upsilon^j = \Upsilon_{\tau_j}^{\text{OvA},n_j}$. Then, for all distributions D and functions $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$,*

$$\mathbf{R}_D^{\ell^H}[\Upsilon^{\text{cas}} \circ \mathbf{f}] \leq \frac{1}{2} \max_{j \in [h]} (\alpha_j + \beta_j) \cdot \mathbf{R}_D^{\psi^{\text{cas}}}[\mathbf{f}]$$

Proof. Let $\mathbf{u} \in \mathbb{R}^d, \mathbf{p} \in \Delta_n$, From Lemmas 8 and 9, we have that

$$\begin{aligned}
\mathbf{R}_{\mathbf{p}}^{\ell^H}[\Upsilon^{\text{cas}}(\mathbf{u})] &\leq \sum_{j=1}^{h} \gamma_j(\mathbf{u}_j) \cdot \mathbf{R}_{A^j(\mathbf{p})}^{\ell^{?,n_j}}[\Upsilon^?_j(\mathbf{u}_j)] \\
&\leq \sum_{j=1}^{h} \gamma_j(\mathbf{u}_j) \left(\frac{\mathbf{1}(\Upsilon^{\text{OvA},n_j}(\mathbf{u}_j) = \perp)}{2(1-\tau_j)} + \frac{\mathbf{1}(\Upsilon^{\text{OvA},n_j}(\mathbf{u}_j) \neq \perp)}{2(1+\tau_j)} \right) \cdot \mathbf{R}_{A^j(\mathbf{p})}^{\psi^{\text{OvA},n_j}}[\mathbf{u}_j] \\
&= \sum_{j=1}^{h} \left(\frac{\beta_j \cdot \mathbf{1}(\Upsilon^{\text{OvA},n_j}(\mathbf{u}_j) = \perp)}{(1-\tau_j)} + \frac{\alpha_j \cdot \mathbf{1}(\Upsilon^{\text{OvA},n_j}(\mathbf{u}_j) \neq \perp)}{(1+\tau_j)} \right) \cdot \mathbf{R}_{A^j(\mathbf{p})}^{\psi^{\text{OvA},n_j}}[\mathbf{u}_j]
\end{aligned}$$

For each $j \in [h]$, the coefficients of both the terms within parantheses (i.e. $\frac{\alpha_j}{1+\tau_j}$ and $\frac{\beta_j}{1-\tau_j}$) both evaluate to $\frac{\alpha_j+\beta_j}{2}$ when the thresholds τ_j are set as $\tau_j = \frac{\alpha_j-\beta_j}{\alpha_j+\beta_j}$. In fact it can easily be seen that this value of τ_j minimizes the worst-case coefficient of $\mathbf{R}^{\psi^{\text{OvA},n_j}}_{A^j(\mathbf{p})}[\mathbf{u}_j]$ in the bound. Thus, we have

$$\begin{aligned}
\mathbf{R}^{\ell^H}_{\mathbf{p}}[\Upsilon^{\text{cas}}(\mathbf{u})] &\leq \sum_{j=1}^{h} \frac{1}{2}(\alpha_j + \beta_j) \cdot \mathbf{R}^{\psi^{\text{OvA},n_j}}_{A^j(\mathbf{p})}[\mathbf{u}_j] \\
&\leq \frac{1}{2} \max_{j\in[h]}(\alpha_j + \beta_j) \cdot \sum_{j=1}^{h} \mathbf{R}^{\psi^{\text{OvA},n_j}}_{A^j(\mathbf{p})}[\mathbf{u}_j] \\
&= \frac{1}{2} \max_{j\in[h]}(\alpha_j + \beta_j) \cdot \mathbf{R}^{\psi^{\text{cas}}}_{\mathbf{p}}[\mathbf{u}] \,.
\end{aligned}$$

The Theorem now follows from linearity of expectation. □