Supplementary material: Structured Output Learning with Abstention : application to Accurate Opinion Prediction

1 Proof of theorem 1

We aim at minimizing the risk of predictor (h, r) based on an estimate $\hat{g}$ of the conditional density $\mathbb{E}_{y|x}\psi_{wa}(y)$:

$$(h(x), r(x)) = \underset{(y_h, y_r)\in\mathcal{Y}^{H,R}}{\arg\min} \langle C\psi_a(y_h, y_r), \hat{g}(x)\rangle,$$

and the corresponding risk is given by :

$$\mathcal{R}(h, r) = \mathbb{E}_x\langle C\psi_a(h(x), r(x)), \mathbb{E}_{y|x}\psi_{wa}(y)\rangle.$$

The optimal predictor $(h^\star, r^\star)$ is the one which is based on the estimate $\hat{g} = \mathbb{E}_{y|x}\psi_{wa}(y)$ which minimized the surrogate risk $\mathcal{L}$:

$$h^\star(x), r^\star(x) = \underset{(y_h, y_r)\in\mathcal{Y}^{H,R}}{\arg\min} \langle C\psi_a(y_h, y_r), \mathbb{E}_{y|x}\psi_{wa}(y)\rangle,$$

and the corresponding risk of the optimal predictor is :

$$\mathcal{R}(h^*, r^*) = \mathbb{E}_x\langle C\psi_a(h^*(x), r^*(x)), \mathbb{E}_{y|x}\psi_{wa}(y)\rangle.$$

Suppose that we have first solved the learning step and we have computed an estimate $\hat{g}(x)$, we have :

$$\begin{aligned}\mathcal{R}(h, r) - \mathcal{R}(h^\star, r^\star) &= \mathbb{E}_x\langle C[\psi_a(h(x), r(x)) - \psi_a(h^\star(x), r^\star(x))], \mathbb{E}_{y|x}\psi_{wa}(y)\rangle \\ &= \mathbb{E}_x\langle C\psi_a(h(x), r(x))(\mathbb{E}_{y|x}[\psi_{wa}(y)] - \hat{g}(x))\rangle \\ &+\mathbb{E}_x\langle C\psi_a(h(x), r(x)), \hat{g}(x)\rangle \\ &-\mathbb{E}_x\langle C\psi_a(h^\star(x), r^\star(x)), \mathbb{E}_{y|x}\psi_{wa}(y)\rangle.\end{aligned}$$

The first term can be bounded by taking the supremum over $\mathcal{Y}^{H,R}$ of the possible predictions :

$$\begin{aligned}&\mathbb{E}_x\langle C\psi_a(h(x), r(x)), (\mathbb{E}_{y|x}[\psi_{wa}(y)] - \hat{g}(x))\rangle \\ \leq\ & \mathbb{E}_x\Big(\sup_{(y_h, y_r)\in\mathcal{Y}^{H,R}} |\langle C\psi_a(y_h, y_r), (\hat{g}(x) - \mathbb{E}_{y|x}[\psi_{wa}(y)])\rangle|\Big).\end{aligned}$$

The second and third term can be rewritten using the definition of the predictors :

$$\langle C\psi_a(h(x), r(x)), \hat{g}(x)\rangle = \inf_{(y_h,y_r)\in\mathcal{Y}^{H,R}} \langle C\psi_a(y_h, y_r), \hat{g}(x)\rangle$$
$$\langle C\psi_a(h^\star(x), r^\star(x)), \mathbb{E}_{y|x}\psi_{wa}(y)\rangle = \inf_{(y_h,y_r)\in\mathcal{Y}^{H,R}} \langle C\psi_a(y_h, y_r), E_{y|x}\psi_{wa}(y)\rangle.$$

The two terms can then be combined :

$$\begin{aligned}\inf_{(y_h,y_r)\in\mathcal{Y}^{H,R}} \langle C\psi_a(y_h, y_r), \hat{g}(x)\rangle - \inf_{(y_h,y_r)\in\mathcal{Y}^{H,R}} \langle C\psi_a(y_h, y_r), E_{y|x}\psi_{wa}(y)\rangle \\ \leq \sup_{(y_h,y_r)\in\mathcal{Y}^{H,R}} |\langle C\psi_a(y_h, y_r), (\hat{g}(x) - E_{y|x}\psi_{wa}(y))\rangle|.\end{aligned}$$

Which gives the same term as above. By combining the results :

$$\begin{aligned}\mathcal{R}(h, r) - \mathcal{R}(h^\star, r^\star) &\leq 2\mathbb{E}_x\left(\sup_{(y_h,y_r)\in\mathcal{Y}^{H,R}} |\langle C\psi_a(y_h, y_r), (\hat{g}(x) - E_{y|x}\psi_{wa}(y))\rangle|\right) \\ &\leq 2\mathbb{E}_x\left(\sup_{(y_h,y_r)\in\mathcal{Y}^{H,R}} \|C\psi_a(y_h, y_r)\|_{\mathbb{R}^q} \|(\hat{g}(x) - E_{y|x}\psi_{wa}(y))\|_{\mathbb{R}^q}\right) \\ &\leq 2 \sup_{(y_h,y_r)\in\mathcal{Y}^{H,R}} \|\psi_a(y_h, y_r)\|_{\mathbb{R}_p} \cdot \|C\| \cdot \mathbb{E}_x\left(\|(\hat{g}(x) - E_{y|x}\psi_{wa}(y))\|_{\mathbb{R}^q}\right) \\ &\leq 2 \sup_{(y_h,y_r)\in\mathcal{Y}^{H,R}} \|\psi_a(y_h, y_r)\|_{\mathbb{R}_p} \cdot \|C\| \cdot \sqrt{\mathbb{E}_x\left(\|(\hat{g}(x) - E_{y|x}\psi_{wa}(y))\|^2_{\mathbb{R}^q}\right)}.\end{aligned}$$

Where $\|C\| = \sup_{x\in\mathbb{R}^p | \|x\|\leq 1} \|Cx\|_{\mathbb{R}^q}$ is the operator norm and the last line is obtained using Jensen inequality.

Finally we expand the form under the square root :

$$\begin{aligned}\mathbb{E}_x[\|(\hat{g}(x) - E_{y|x}\psi_{wa}(y))\|^2_{\mathbb{R}^q}] &= \mathbb{E}_x\|\hat{g}(x)\|^2_{\mathbb{R}^q} + \|E_{y|x}\psi_{wa}(y))\|^2_{\mathbb{R}^q} - 2\langle\hat{g}(x), E_{y|x}\psi_{wa}(y)\rangle \\ &= \mathbb{E}_x\|\hat{g}(x)\|^2_{\mathbb{R}^q} - \|E_{y|x}\psi_{wa}(y)\|^2_{\mathbb{R}^q} + 2\langle E_{y|x}\psi_{wa}(y), E_{y|x}\psi_{wa}(y)\rangle \\ &- 2\langle\hat{g}(x), E_{y|x}\psi_{wa}(y)\rangle + \mathbb{E}_{x,y}\|\psi_{wa}(y)\|^2_{\mathbb{R}^q} - \mathbb{E}_{x,y}\|\psi_{wa}(y)\|^2_{\mathbb{R}^q} \\ &= \mathbb{E}_x\|\hat{g}(x)\|^2_{\mathbb{R}^q} + \mathbb{E}_{x,y}\|\psi_{wa}(y)\|^2_{\mathbb{R}^q} - 2\mathbb{E}_{x,y}\langle\hat{g}(x), \psi_{wa}(y)\rangle \\ &- \left(\|E_{y|x}\psi_{wa}(y)\|^2_{\mathbb{R}^q} + \|\psi_{wa}(y)\|^2_{\mathbb{R}^q} - 2\mathbb{E}_{x,y}\langle\|E_{y|x}\psi_{wa}(y), \psi_{wa}(y)\rangle\right) \\ &= \mathbb{E}_{x,y}\|\hat{g}(x) - \psi_{wa}(y)\|^2_{\mathbb{R}^q} - \mathbb{E}_{x,y}\|E_{y|x}\psi_{wa}(y) - \psi_{wa}(y)\|^2_{\mathbb{R}^q}.\end{aligned}$$

Which is equal to $\mathcal{L}(\hat{g}) - \mathcal{L}(\mathbb{E}_{y|x}\psi_{wa})$.

2 Canonical form for some examples of the abstention aware loss

2.1 Canonical form for the Δ_{bin} loss

Let us consider the binary classification with a reject option loss :

$$\Delta_a^{bin}(h(x), r(x), y) = \begin{cases} 1 \text{ if } y \neq h(x) \text{ and } r(x) = 1 \\ 0 \text{ if } y = h(x) \text{ and } r(x) = 1 \\ c \text{ if } r(x) = 0 \end{cases},$$

It can also be rewritten as a function of the binary variables :

$$\begin{aligned} \Delta_a^{bin}(h(x), r(x), y) &= r(x)[1 - (h(x) - y)^2] + (1 - r(x))c \\ &= r(x)[1 - h(x) - y + 2h(x)y] + (1 - r(x))c \\ &= y(h(x)r(x)) + (1 - y)(1 - h(x))r(x) + (y + (1 - y))c(1 - r(x)). \end{aligned}$$

Which corresponds to the parameterization proposed in the article.

2.2 Canonical form for the Δ_H loss

Let us consider the hierarchical loss :

$$\Delta_H(h(x), r(x), y) = \sum_{i=1}^{d} c_i 1_{h(x)_i \neq y_i} 1_{h(x)_{p(i)} = y_{p(i)}}.$$

It is defined on objects that respect the hierarchical condition :

$$\forall i \in \{1, \ldots, d\}, \forall y \in \{0,1\}^d \; y_i \leq y_{p(i)},$$

under the hypothesis of a binary vector, the loss can be rewritten :

$$\begin{aligned} \Delta_H(h(x), r(x), y) &= \sum_{i=1}^{d} c_i (h(x)_i - y_i)^2 (1 - (h(x)_{p(i)} - y_{p(i)})^2 \\ &= \sum_{i=1}^{d} c_i (h(x)_i + y_i - 2h(x)_i y_i)(1 - h(x)_{p(i)} - y_{p(i)} + 2h(x)_{p(i)} y_{p(i)}). \end{aligned}$$

Where the second line has been obtained using the fact that for binary variables, $e = e^2$. Due to the hierarchical constraint, we also have $y_i y_{p(i)} = y_i$ and $h(x)_i h(x)_{p(i)} = h(x)_i$:

$$\Delta_H(h(x), r(x), y) = \sum_{i=1}^{d} c_i (h(x)_i (y_{p(i)} - 2y_i) + h(x)_{p(i)} y_i).$$

Which corresponds to the parameterization proposed in the article.

2.3 Canonical form for the Δ_{Ha} loss

See section 4 of the supplementary material.

3 Proof of theorem 2

Let us recall the problem to solve :

$$\underset{(y_h, y_r) \in \mathcal{Y}^{H,R}}{\arg\min} \langle \psi_a(y_h, y_r, \psi_x \rangle,$$

Using the additional hypothesis over ψ_a we obtain the problem :

$$\hat{h}(x), \hat{r}(x) = \underset{(y_h, y_r) \in \mathcal{Y}^{H,R}}{\arg\min} (y_h^T, y_r^T, (y_h \otimes y_r)^T) M^T \psi_x.$$

Where $\otimes$ is the Kronecker product between 2 vectors. This problem can be transformed into the constrained optimization problem :

$$\hat{h}(x), \hat{r}(x) = \underset{(y_h, y_r) \in \mathcal{Y}^{H,R}}{\arg\min} (y_h^T, y_r^T, c^T) M^T \psi_x.$$
$$\text{s.t. } \big(c = y_h \otimes y_r\big)$$

Let us show that the constraint $c = y_h \otimes y_r$ can be replaced by a set of linear constraints when $h(x)$ and $r(x)$ are two binary vectors:

3.1 Constraints on the c vector

The linearisation of the constraint relies on the following result :

Proposition 1. *Let x and y be 2 binary variables and e the binary variables defined by the formula $e = x \cdot y$ where $\cdot$ denotes the logical AND : $e = 1$ if $x = 1$ and $y = 1$ and 0 else. Then the following holds :*

$$e = x \cdot y \iff \begin{cases} e \leq x \\ e \leq y \\ e \geq x + y - 1 \\ e \geq 0 \end{cases} . \tag{1}$$

This representation can be used to rewrite the constraints on the c vector. By definition of the Kronecker product : $y_h \otimes y_r = \begin{pmatrix} y_{h,1} y_r \\ y_{h,2} y_r \\ \cdot \\ y_{h,d} y_r \end{pmatrix}$ where $y_{h,i}$ is the i_{th} component of y_h.

We write each inequality of (1) as a linear matrix inequality :

$$
\begin{aligned}
&c \leq A_{h,1}y_h\\
&c \leq A_{r,1}y_r\\
&c \geq A_{h,2}y_h + A_{r,2}y_r + b_1\\
&c \geq 0.
\end{aligned}
$$

All these inequality can be merged in a single one :

$$A_{\text{constraints c}} \begin{pmatrix} y_h \\ y_r \\ c \end{pmatrix} \leq b_{\text{constraints c}},$$

where $A_{\text{constraints c}} = \begin{pmatrix}
-I_d & 0_d & I_d & 0_d & 0_d & \cdots & 0_d \\
-I_d & 0_d & 0_d & I_d & 0_d & \cdots & 0_d \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
-I_d & 0_d & \cdots & 0_d & \cdots & \cdots & I_d \\
0_d & -V_1 & I_d & 0_d & 0_d & \cdots & 0_d \\
0_d & -V_2 & 0_d & I_d & 0_d & \cdots & 0_d \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0_d & -V_d & \cdots & 0_d & \cdots & \cdots & I_d \\
I_d & V_1 & -I_d & 0_d & 0_d & \cdots & 0_d \\
I_d & V_2 & 0_d & -I_d & 0_d & \cdots & 0_d \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
I_d & V_d & \cdots & 0_d & \cdots & \cdots & -I_d \\
0_d & 0_d & I_d & 0_d & \cdots & \cdots & \cdots \\
0_d & 0_d & 0_d & I_d & 0_d & \cdots & \cdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0_d & \ddots & \ddots & \ddots & \ddots & 0_d & I_d
\end{pmatrix}$

and $b_{\text{constraints c}} = \begin{pmatrix} 0_{d^2,1} \\ 0_{d^2,1} \\ 1_{d^2,1} \\ 0_{d^2,1} \end{pmatrix}$. I_d is the $d \times d$ identity matrix, 0_d the $d \times d$ matrix full of 0, $0_{d^2,1}$ the d^2 dimensional vector full of 0 and $1_{d^2,1}$ the d^2 dimensional vector full of 1.

V_i is the $d \times d$ matrix such that all its entries are 0 except the i^{th} which is 1. The 4 distinct blocks correspond to the 4 different constraints given in 1.

4 Construction of the linear program for the Hierarchical loss with abstention

Let us suppose that our prediction are the assignments of a d nodes binary tree with an abstention label a.

We recall the parameterization of our loss :

$$\Delta_{Ha}(h(x), r(x), y) = \sum_{i=1}^{d} c_{Ai} 1_{\{f_i^{h,r}=a, f_{p(i)}^{h,r}=y_{p(i)}\}}$$
$$+ c_{A_c i} 1_{\{f_i^{h,r}\neq y_i, f_{p(i)}^{h,r}=a\}}$$
$$+ c_i 1_{\{f_i^{h,r}\neq y_i, f_{p(i)}^{h,r}=y_{p(i)}, f_i^{h,r}\neq a\}}.$$

With $f^{h,r}$ a prediction function built from the pair $(h,r) : \mathcal{X} \to \mathcal{Y}^{H,R}$:

$$f^{h,r}(x)^T = [f_1^{h,r}(x), \dots, f_d^{h,r}(x)],$$
$$f_i^{h,r}(x) = 1_{h(x)_i=1} 1_{r(x)_i=1} + a 1_{r(x)_i=0},$$

In what follows, we denote by $p(i)$ the index of the parent of the i according to the underlying tree and suppose that our trees are rooted at the node of index 0 for which the label is 1 and there is no abstention.

We recall the set of constraints we used to define $\mathcal{Y}^{H,R}$ for the Ha loss :

- Abstention at 2 consecutive nodes is forbidden : $\forall i \in \{1, \dots, d\}$ $r(x)_i + r(x)_{p(i)} \leq 1$.
- A node can be set to one only if its parent is set to 1 or if the predictor abstained itself from predicting it : $h(x)_i r(x)_{p(i)} \leq h(x)_{p(i)} r(x)_{p(i)}$.

Since $h(x)$ and $r(x)$ are both binary vectors, one can rewrite the loss as a function of these predictions :

$$\Delta_{Ha}(h(x), r(x), y) = \sum_{i=1}^{n} c_i (h(x)_i - y_i)^2 [1 - (h(x)_{p(i)} - y_{p(i)})^2] r(x)_i r(x)_{p(i)}$$
$$+ c_{Ai}(1 - r(x)_i)[1 - (h(x)_{p(i)} - y_{p(i)})^2]$$
$$+ c_{A_c i}(h(x)_i - y_i)^2 (1 - r(x)_{p(i)}).$$

We develop and simplify according to the fact that for any binary variable b, we have $b^2 = b$:

$$\Delta_{Ha}(h(x), r(x), y) = \sum_{i=1}^{n} c_i (h(x)_i + y_i - 2h(x)_i y_i)$$
$$[1 - (h(x)_{p(i)} + y_{p(i)} - 2h(x)_{p(i)} y_{p(i)})] r(x)_i r(x)_{p(i)}$$
$$+ c_{Ai}(1 - r(x)_i)[1 - (h(x)_{p(i)} + y_{p(i)} - 2h(x)_{p(i)} y_{p(i)})]$$
$$+ c_{A_c i}(h(x)_i + y_i - 2h(x)_i y_i)(1 - r(x)_{p(i)}).$$

We take into account the known constraints :

- The hierarchical constraint can be written : $(1 - h(x)_{p(i)}) r(x)_{p(i)} = 1 \implies h(x)_i = 0$ which leads to the equality : $(1 - h(x)_{p(i)}) r(x)_{p(i)} h(x)_i = 0 \iff h(x)_{p(i)} h(x)_i r(x)_{p(i)} = h(x)_i r(x)_{p(i)}$.

- The non consecutive abstention constraint implies $r(x)_i r(x)_{p(i)} = r(x)_i + r(x)_{p(i)} - 1$.

We treat the 3 terms of the l_{HA} loss separately as follows :

$$\Delta_{Ha}(h(x), r(x), y) = \sum_{i=1}^{n} c_i A_i(x) + c_{Ai} B_i(x) + c_{A_c i} C_i(x).$$

And rewrite each of this term as a linear combination of the unknown variables (corresponding to some elements of the vector $\begin{pmatrix} h(x) \\ r(x) \\ h(x) \otimes r(x) \end{pmatrix}$):

First term :

$$\begin{aligned}
A_i(x) &= (h(x)_i + y_i - 2h(x)_i y_i)(1 - h(x)_{p(i)} - y_{p(i)} + 2h(x)_{p(i)} y_{p(i)}) r(x)_i r(x)_{p(i)} \\
&= (h(x)_i (1 - 2y_i) + y_i)(h(x)_{p(i)}(2y_{p(i)} - 1) + 1 - y_{p(i)}) r(x)_i r(x)_{p(i)} \\
&= \Big(h(x)_i h(x)_{p(i)} (1 - 2y_i)(2y_{p(i)} - 1) + \\
&\quad h(x)_i (1 - y_{p(i)})(1 - 2y_i) + h(x)_{p(i)} y_i (2y_{p(i)} - 1) + y_i (1 - y_{p(i)}) \Big) r(x)_i r(x)_{p(i)} \\
&= h(x)_i h(x)_{p(i)} r(x)_{p(i)} r(x)_i (1 - 2y_i)(2y_{p(i)} - 1) + \\
&\quad h(x)_i r(x)_i r(x)_{p(i)} (1 - y_{p(i)})(1 - 2y_i) + \\
&\quad h(x)_{p(i)} r(x)_i r(x)_{p(i)} y_i (2y_{p(i)} - 1) + \\
&\quad r(x)_i r(x)_{p(i)} y_i (1 - y_{p(i)}).
\end{aligned}$$

Using the first constraint, we have : $h(x)_i h(x)_{p(i)} r(x)_{p(i)} r(x)_i = h(x)_i r(x)_{p(i)} r(x)_i$. Using this reduction and the second constraint we obtain the equation :

$$
\begin{aligned}
A_i(x) = h(x)_i r(x)_i \Big((1-2y_i)(2y_{p(i)}-1) + (1-y_{p(i)})(1-2y_i) \Big) + \\
h(x)_i r(x)_{p(i)} \Big((1-2y_i)(2y_{p(i)}-1) + (1-y_{p(i)})(1-2y_i) \Big) + \\
h(x)_{p(i)} r(x)_i \Big(y_i(2y_{p(i)}-1) \Big) + \\
h(x)_{p(i)} r(x)_{p(i)} \Big(y_i(2y_{p(i)}-1) \Big) + \\
h(x)_i \Big(-(1-2y_i)(2y_{p(i)}-1) - (1-y_{p(i)})(1-2y_i) \Big) + \\
h(x)_{p(i)} \Big(y_i(1-2y_{p(i)}) \Big) + \\
r(x)_i \Big(y_i(1-y_{p(i)}) \Big) + \\
r(x)_{p(i)} \Big(y_i(1-y_{p(i)}) \Big) + \\
\Big(y_i(y_{p(i)}-1) \Big).
\end{aligned}
$$

Second term :

$$
\begin{aligned}
B_i(x) &= (1-r(x)_i)(1-h(x)_{p(i)} - y_{p(i)} + 2h(x)_{p(i)} y_{p(i)}) \\
&= h(x)_{p(i)} r(x)_i \Big(1-2y_{p(i)} \Big) + \\
&h(x)_{p(i)} \Big(2y_{p(i)}-1 \Big) + \\
&r(x)_i \Big(y_{p(i)}-1 \Big) + \\
&\Big(1-y_{p(i)} \Big).
\end{aligned}
$$

Third term :

$$
\begin{aligned}
C_i(x) &= h(x)_i + y_i - 2h(x)_i y_i)(1-r(x)_{p(i)}) \\
&= h(x)_i r(x)_{p(i)} \Big(2y_i - 1 \Big) + \\
&h(x)_i \Big(1-2y_i \Big) + \\
&r(x)_{p(i)} \Big(-y_i \Big) + \\
&\Big(y_i \Big).
\end{aligned}
$$

Sum of the three terms

Based on the previous results we express the loss as a linear combination of the different variables previously expressed :

$$\Delta_{Ha}(h(x), r(x), y) = \Big(\sum_{i=1}^{n} a_{(i)}^{(1)} h(x)_i + a_{(i)}^{(2)} h(x)_i r(x)_{p(i)} + a_{(i)}^{(3)} h(x)_{p(i)} r(x)_i + a_{(i)}^{(4)} h(x)_i r(x)_i + a_{(i)}^{(5)} r(x)_i +$$
$$a_{(i)}^{(6)} h(x)_{p(i)} + a_{(i)}^{(7)} r(x)_{p(i)} + a_{(i)}^{(8)} h(x)_{p(i)} r(x)_{p(i)} + a_{(i)}^{(9)}\Big).$$

With the following table of correspondency $\forall k \in \{1, \ldots, d\}$:

$$\begin{aligned}
a_{(i)}^{(1)} &= -c_i((1-2y_i)(2y_{p(i)}-1) + (1-y_{p(i)})(1-2y_i)) + c_{A_c i}(1-2y_i) \\
a_{(i)}^{(2)} &= c_i((1-2y_i)(2y_{p(i)}-1) + (1-y_{p(i)})(1-2y_i)) + c_{A_c i}(2y_i-1) \\
a_{(i)}^{(3)} &= c_i(y_i(2y_{p(i)}-1)) + c_{Ai}(1-2y_{p(i)}) \\
a_{(i)}^{(4)} &= c_i((1-2y_i)(2y_{p(i)}-1) + (1-y_{p(i)})(1-2y_i)) \\
a_{(i)}^{(5)} &= c_i y_i(1-y_{p(i)}) + c_{Ai}(y_{p(i)}-1) \\
a_{(i)}^{(6)} &= c_i y_i(1-2y_{p(i)}) + c_{Ai}(2y_{p(i)}-1) \\
a_{(i)}^{(7)} &= c_i y_i(1-y_{p(i)}) - c_{A_c i} y_i \\
a_{(i)}^{(8)} &= c_i y_i(2y_{p(i)}-1) \\
a_{(i)}^{(9)} &= c_i y_i(y_{p(i)}-1) + c_{Ai}(1-y_{p(i)}) + c_{A_c i} y_i.
\end{aligned}$$

We introduce a new vector of variables $g = \begin{pmatrix} g^{(1)} \\ g^{(2)} \\ \vdots \\ g^{(8)} \end{pmatrix}$ where each of the n dimensional vectors $g^{(k)}$ is defined as follows : $\forall i \in \{1, \ldots, n\}$

$$\begin{aligned}
g_i^{(1)} &= h_i \\
g_i^{(2)} &= h_i r_{p_i} \\
g_i^{(3)} &= h_{p_i} r_i \\
g_i^{(4)} &= h_i r_i \\
g_i^{(5)} &= r_i \\
g_i^{(6)} &= h_{p_i} \\
g_i^{(7)} &= r_{p_i} \\
g_i^{(8)} &= h_{p_i} r_{p_i}.
\end{aligned}$$

The last variables are redundant since g_{p_i} and g_i are the same except at the root and leaves. Let us denote by A_h the adjacency matrix of the underlying hierarchy and $\forall p \in \{1, \ldots, 8\}$ $y^{(p)} = \begin{pmatrix} y_1^{(p)} \\ \cdot \\ y_d^{(p)} \end{pmatrix}$ and $a_{\bar{(p)}} = \begin{pmatrix} a_{\bar{(p)}1} \\ \cdot \\ a_{\bar{(p)}d} \end{pmatrix}$. Then we have

$$\begin{aligned} y^{(6)} &= A_h y^{(1)} \\ y^{(7)} &= A_h y^{(5)} \\ y^{(8)} &= A_h y^{(4)}. \end{aligned}$$

Let us denote by $a^{(p)} = \begin{pmatrix} a_1^{(p)} \\ a_2^{(p)} \\ \vdots \\ a_n^{(p)} \end{pmatrix}$, on can rewrite the loss $l(y^{(A)}, y)$ using the reduced set of variables :

$$\Delta_{Ha}(h(x), r(x), y) = \sum_{p=1}^{5} \left((a^{(p)})^T g^{(p)} \right) + (a^{(6)})^T A_h g^{(1)} + (a^{(7)})^T A_h y^{(5)} + (a^{(8)})^T A_h y^{(4)}.$$

This is a linear program by choosing the cost vector c and the variable g' :

$$c = \begin{pmatrix} a^{(1)} + A_h^T a^{(6)} \\ a^{(2)} \\ a^{(3)} \\ a^{(4)} + A_h^T a^{(8)} \\ a^{(5)} + A_h^T a^{(7)} \end{pmatrix} \quad g' = \begin{pmatrix} g^{(1)} \\ g^{(2)} \\ g^{(3)} \\ g^{(4)} \\ g^{(5)}. \end{pmatrix}$$

Leading to the reduced form :

$$l(y^{(A)}, y) = c^T g'.$$

In our applications, the abstention aware predictor we built relied on solving problems of the form :

$$\arg\min_{y^{(A)}} \sum_{k=1}^{N} \alpha_k(x) \Delta_{Ha}(h(x), r(x), y_k).$$

Where (x_k, y_k) $k \in \{1, \ldots, N\}$ are labelled example of a N sample training set and $(x, f^{h,r})$ correspond to the new input x for which we look for the best prediction $f^{h,r}$.

According to the previous results, we denote by c_k the cost vector computed from the term $l(y^{(A)}, y_k)$ and $\bar{c}(x) = \sum_{k=1}^{n} \alpha_k(x) c_k$ the full cost vector of the previous minimization problem. The minimization problem can be rewritten explicit in terms of the vector of variables g' by making the constraints between its different parts explicit :

$$\begin{aligned}
\arg\min_{y^{(A)}} \sum_{k=1}^{N} \alpha_k(x)\Delta_{Ha}(h(x), r(x), y_k) = \quad & \arg\min_{g' \in \{0,1\}^{8n}} \; c^T g' \\
\text{subject to} \quad & g^{(2)} = g^{(1)} \odot A_h g^{(5)}, \\
& g^{(3)} = A_h g^{(1)} \odot g^{(5)}, \\
& g^{(4)} = g^{(1)} \odot g^{(5)}, \\
& g^{(2)} \leq A_h g^{(4)}, \\
& g^{(5)} \in \mathcal{Y}_r.
\end{aligned}$$

Where $\mathcal{Y}_r$ is the space of d dimensional binary vectors such that $\forall y \in \mathcal{Y}_r \; \forall i \in \{1, \dots, d\} \; y_i + y_{p(i)} \leq 1$. The 3 first constraints are given by construction of the g' vector from 2 underlying vectors $r(x)$ and $h(x)$. The fourth line is the generalized hierarchical constraint : $\forall i \in 1, \dots, n \; h(x)_i r(x)_{p(i)} \leq h(x)_{p(i)} r(x)_{p(i)}$. The fifth line corresponds to the hypothesis of no 2 consecutive abstentions.

We turn this program into a canonical linear program with binary value constraints :

$$\begin{aligned}
\arg\min_{g} \mathcal{L}(g) = \quad & \arg\min_{g' \in \{0,1\}^{8n}} \; c^T g' \\
\text{subject to} \quad & g^{(2)} \leq g^{(1)}, \\
& g^{(2)} \leq A_h g^{(5)}, \\
& g^{(2)} \geq g^{(1)} + A_h g^{(5)} - 1, \\
& g^{(3)} \leq A_h g^{(1)}, \\
& g^{(3)} \leq g^{(5)}, \\
& g^{(3)} \geq A_h g^{(1)} + g^{(5)} - 1, \\
& g^{(4)} \leq g^{(1)}, \\
& g^{(4)} \leq g^{(5)}, \\
& g^{(4)} \geq g^{(1)} + g^{(5)} - 1, \\
& g^{(2)} \leq A_h g^{(4)}, \\
& I_d + A_h g^{(5)} \leq 1.
\end{aligned}$$

In our experiments, this integer linear program is solved using the python cylp binder to the Cbc library and directly implemented using sparse representations.

4.1 Hierarchical classification of MRI images

The Medical Retrieval Task of the ImageCLEF 2007 challenge provided a set of medical images aligned with a code corresponding to a class in a predefined hierarchy. A class is described by 4 values encoded as follows :

- T (Technical) : image modality
- D (Directional) : body orientation
- A (Anatomical) : body region examined
- B (Biological) : biological system examined

In our experiments we focus on the D and A tasks and reuse the representation proposed in [DKLD08] and freely available at the page : `http://ijs.si/DragiKocev/PhD/resources/doku.php?id=hmc_classification`. Each dataset contains an existing train test split with 10000 labeled objects for training and 1006 for testing. The A task consist in predicting the assignment of a 96 nodes binary tree of maximal depth 3 (an example of label at depth 3 is : upper extremity / arm → hand → finger). The D task consist in predicting the assignment of a 46 nodes binary tree of maximal depth 3 (an example of label at depth 3 is : sagittal → lateral, right-left → inspiration). The complete hierarchy is described in [LSK+03]

The table below contains the results in terms of Hamming Loss for the problem of hierarchical classification.

Method	Hamming loss
H Regression	0.0189
Depth weighted Regression	0.0193
Uniform Regression	0.0218
Binary SVC	0.0197

Table 1: Results on the ImageCLEF2007d task

Method	Hamming loss
H Regression	0.0065
Depth weighted Regression	0.0068
Uniform Regression	0.0102
Binary SVC	0.0071

Table 2: Results on the ImageCLEF2007a task

We compare our method (H regression) using the sibbling weighted scheme described in the article against our same method (Uniform regression) with a uniform weighted scheme ($c_i = 1 \ \forall i \in \{1, \dots, d\}$), a depth weighted scheme ($c_i = \frac{c_{p(i)}}{N_d} \ \forall i \in \{1, \dots, d\}$ where N_d is the number of nodes at depth d i.e. separated from the root by $d+1$ nodes) and against the binary relevance Support Vector Classifier approach (binary SVC) which consist in training one SVM classifier for each node and applying the Hierarchical condition in a second time by switching to 0 all the nodes which for which the parent node has the label 0. We used the gaussian kernel for the input data in all 3 methods and tuned the hyperparameters by 5 folds cross validation and report the results on the available test set.

These results illustrate the choice of the sibbling weighted scheme for the H loss since it retrieve the best results. Moreover, taking the structured representation into account is shown to improve the results over the Binary SVC approach on both tasks.

References

[DKLD08] Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Sašo Džeroski. Hierchical annotation of medical images. In *Proceedings of the 11th International Multiconference - Information Society IS 2008*, pages 174–181. IJS, Ljubljana, 2008.

[LSK+03] Thomas Martin Lehmann, Henning Schubert, Daniel Keysers, Michael Kohnen, and Berthold B Wein. The irma code for unique classification of medical images. In *Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*, volume 5033, pages 440–452. International Society for Optics and Photonics, 2003.