ADIOS: Architectures Deep In Output Space
Supplementary Material

1. Proofs

Proposition 1 *Let $\mathcal{D} = \{(x, y)\}_{1 \leq i \leq m}$ sampled from the joint distribution $(X \times Y)$, $((X, G_1), G_2)$ a Markov Blanket Chained partition of this joint distribution, and $G_i(y_k)$ denotes the projection of a set of labels y_k to G_i. The CE loss of a hypothesis h from a hypothesis class H is given by:*

$$\ell(h; \mathcal{D}) = -\sum_{i=1}^{m} \underbrace{G_2(y_i) \cdot \log(h(G_1(y_i), x_i)) + (1 - G_2(y_i)) \cdot \log(1 - h(G_1(y_i), x_i))}_{\textit{second layer cross-entropy } \mathcal{L}_2(h, x_i)} \tag{1}$$

$$+ \underbrace{G_1(y_i) \cdot \log(h(x_i)) + (1 - G_1(y_i)) \cdot \log(1 - h(x_i))}_{\textit{first layer cross-entropy } \mathcal{L}_1(h, x_i)} \tag{2}$$

Proof:

$$\begin{aligned} P(\mathcal{D}|h) &= \prod_{i=1}^{m} P(x_i, y_i|h) = \prod_{i=1}^{m} P(y_i|h, x_i) \cdot P(x_i) && (x_i \text{ is independent of } h) \\ &= \prod_{i=1}^{m} P(G_1(y_i), G_2(y_i)|h, x_i) \cdot P(x_i) && (y_i = (G_1(y_i), G_2(y_i))) \\ &= \prod_{i=1}^{m} P(G_2(y_i)|h, G_1(y_i), x_i) \cdot P(G_1(y_i)|h, x_i) \cdot P(x_i) && (\text{product rule}) \end{aligned} \tag{3}$$

Denoting $g_{2_{ij}}$ (resp. $g_{1_{ij}}$) the j-th component of the vector $G_2(y_i)$ (resp. $G_1(y_i)$), the corresponding conditional probabilities can be expressed as:

$$P(g_{2_{ij}}|h, g_{1_i}, x_i) = \begin{cases} h(g_{1_i}) & \text{if } g_{2_{ij}, x_i} = 1 \\ 1 - h(g_{1_i}, x_i) & \text{if } g_{2_{ij}} = 0 \end{cases} \qquad P(g_{1_{ik}}|h, x_i) = \begin{cases} h(x_i) & \text{if } g_{1_{ik}} = 1 \\ 1 - h(x_i) & \text{if } g_{1_{ik}} = 0 \end{cases}$$

Therefore, since the labels belonging to the same layer are independent given those of the previous layer due to the Markov Blanket partitioning, the probability of the data given the hypothesis is:

$$P(\mathcal{D}|h) = \prod_{i=1}^{m} \left[\Big(\prod_{j=1}^{|G_2|} h(g_{1_i}, x_i)^{g_{2_{ij}}} \cdot (1 - h(g_{1_i}, x_i))^{(1 - g_{2_{ij}})} \Big) \cdot \Big(\prod_{k=1}^{|G_1|} h(x_i)^{g_{1_{ik}}} \cdot (1 - h(x_i))^{(1 - g_{1_{ik}})} \Big) \cdot P(x_i) \right]$$

Hence, since $P(x_i)$ is independent from the hypothesis ($\forall i$), the maximum likelihood hypothesis is:

$$P(\mathcal{D}|h) = \prod_{i=1}^{m} \left[\Big(\prod_{j=1}^{|G_2|} h(g_{1_i}, x_i)^{g_{2_{ij}}} \cdot (1 - h(g_{1_i}, x_i))^{(1 - g_{2_{ij}})} \Big) \cdot \Big(\prod_{k=1}^{|G_1|} h(x_i)^{g_{1_{ik}}} \cdot (1 - h(x_i))^{(1 - g_{1_{ik}})} \Big) \right]$$

Taking the negative log-likelihood and using the vectorial notation completes the proof ■