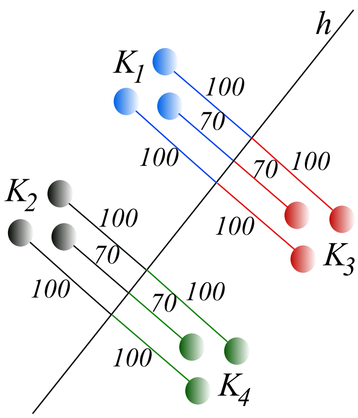
Simultaneous Learning of Trees and Representations for Extreme Classification with Application to Language Modeling (Supplementary material)

9. Geometric interpretation of probabilities $p_j^{(n)}$ and $p_{j|i}^{(n)}$



Discrete:
$p_1^{(n)} = \frac{6}{12} = \mathbf{0.5}$
$p_{1|1}^{(n)} = \frac{3}{3} = \mathbf{1}, \quad p_{1|2}^{(n)} = \frac{3}{3} = \mathbf{1}, \quad p_{1|3}^{(n)} = \frac{0}{3} = \mathbf{0}, \quad p_{1|4}^{(n)} = \frac{0}{3} = \mathbf{0}$

Continuous:
$p_1^{(n)} = \frac{1}{12}(\sigma(100) + \sigma(70) + \ldots + \sigma(-70) + \sigma(-100)) \approx \mathbf{0.5}$
$p_{1|1}^{(n)} = \frac{1}{3}(\sigma(100) + \sigma(70) + \sigma(100)) \approx \mathbf{1}$
$p_{1|2}^{(n)} = \frac{1}{3}(\sigma(100) + \sigma(70) + \sigma(100)) \approx \mathbf{1}$
$p_{1|3}^{(n)} = \frac{1}{3}(\sigma(-100) + \sigma(-70) + \sigma(-100)) \approx \mathbf{0}$
$p_{1|4}^{(n)} = \frac{1}{3}(\sigma(-100) + \sigma(-70) + \sigma(-100)) \approx \mathbf{0}$

Figure 3. The comparison of discrete and continuous definitions of probabilities $p_j^{(n)}$ and $p_{j|i}^{(n)}$ on a simple example with $K = 4$ classes and binary tree ($M = 2$). n is an exemplary node, e.g. root. σ denotes sigmoid function. Color circles denote data points.

Remark 3. *One could define $p_j^{(n)}$ as the ratio of the number of examples that reach node n and are sent to its j^{th} child to the total the number of examples that reach node n and $p_{j|i}^{(n)}$ as the ratio of the number of examples that reach node n, correspond to label i, and are sent to the j^{th} child of node n to the total the number of examples that reach node n and correspond to label i. We instead look at the continuous counter-parts of these discrete definitions as given by Equations 8 and 9 and illustrated in Figure 3 (note that continuous definitions have elegant geometric interpretation based on margins), which simplifies the optimization problem.*

10. Theoretical proofs

Proof of Lemma 1. Recall the form of the objective defined in 6:

$$\begin{aligned} J_n &= \frac{2}{M}\sum_{i=1}^{K} q_i^{(n)}\left(\sum_{j=1}^{M} |p_j^{(n)} - p_{j|i}^{(n)}|\right) \\ &= \frac{2}{M}\mathbb{E}_{i\sim q^{(n)}}\left[f_n^J(i, p_{\cdot|\cdot}^{(n)}, q^{(n)})\right] \end{aligned}$$

Where:

$$\begin{aligned} f_n^J(i, p_{\cdot|\cdot}^{(n)}, q^{(n)}) &= \sum_{j=1}^{M}\left|p_j^{(n)} - p_{j|i}^{(n)}\right| = \sum_{j=1}^{M}\left|p_{j|i}^{(n)} - \sum_{i'=1}^{K} q_{i'}^{(n)} p_{j|i'}^{(n)}\right| \\ &= \sum_{j=1}^{M}\left|\sum_{i'=1}^{K}(\mathbb{1}_{i=i'} - q_{i'}^{(n)}) p_{j|i'}^{(n)}\right| \end{aligned}$$

Hence:

$$\frac{\partial f_n^J(i, p_{.|.}^{(n)}, q^{(n)})}{\partial p_{j|i}^{(n)}} = (1 - q_i^{(n)}) \operatorname{sign}(p_{j|i}^{(n)} - p_j^{(n)})$$

And:

$$\begin{aligned} \frac{\partial f_n^J(i, p_{.|.}^{(n)}, q^{(n)})}{\partial \log p_{j|i}^{(n)}} &= (1 - q_i^{(n)}) \operatorname{sign}(p_{j|i}^{(n)} - p_j^{(n)}) \frac{\partial p_{j|i}^{(n)}}{\partial \log p_{j|i}^{(n)}} \\ &= (1 - q_i^{(n)}) \operatorname{sign}(p_{j|i}^{(n)} - p_j^{(n)}) p_{j|i}^{(n)} \end{aligned}$$

By assigning each label j to a specific child i under the constraint that no child has more than L labels, we take a step in the direction $\partial E \in \{0, 1\}^{M \times K}$, where:

$$\forall i \in [1, K], \quad \textstyle\sum_{j=1}^{M} \partial E_{j,i} = 1$$

and

$$\forall j \in [1, M], \quad \textstyle\sum_{i=1}^{K} \partial E_{j,i} \leq L$$

Thus:

$$\begin{aligned} \frac{\partial J_n}{\partial p_{.|.}^{(n)}} \partial E &= \frac{2}{M} \frac{\mathbb{E}_{i \sim q^{(n)}} \left[f_n^J(i, p_{.|.}^{(n)}, q^{(n)}) \right]}{\partial p_{.|.}^{(n)}} \partial E \\ &= \frac{2}{M} \sum_{i=1}^{K} q_i^{(n)} (1 - q_i^{(n)}) \sum_{j=1}^{M} \left(\operatorname{sign}(p_{j|i}^{(n)} - p_j^{(n)}) \partial E_{j,i} \right) \end{aligned} \tag{13}$$

And:

$$\frac{\partial J_n}{\partial \log p_{.|.}^{(n)}} \partial E = \frac{2}{M} \sum_{i=1}^{K} q_i^{(n)} (1 - q_i^{(n)}) \sum_{j=1}^{M} \left(\operatorname{sign}(p_{j|i}^{(n)} - p_j^{(n)}) p_{j|i}^{(n)} \partial E_{j,i} \right) \tag{14}$$

If there exists such an assignment for which 13 is positive, then the greedy method proposed in 2 finds it. Indeed, suppose that Algorithm 2 assigns label i to child j and i' to j'. Suppose now that another assignment $\partial E'$ sends i to j' and i to j'. Then:

$$\frac{\partial J_n}{\partial p_{.|.}^{(n)}} \left(\partial E - \partial E' \right) = \left(\frac{\partial J_n}{\partial p_{j|i}^{(n)}} + \frac{\partial J_n}{\partial p_{j'|i'}^{(n)}} \right) - \left(\frac{\partial J_n}{\partial p_{j|i'}^{(n)}} + \frac{\partial J_n}{\partial p_{j'|i}^{(n)}} \right) \tag{15}$$

Since the algorithm assigns children by descending order of $\frac{\partial J_n}{\partial p_{j|i}^{(n)}}$ until a child j is full, we have:

$$\frac{\partial J_n}{\partial p_{j|i}^{(n)}} \geq \frac{\partial J_n}{\partial p_{j|i'}^{(n)}} \quad \text{and} \quad \frac{\partial J_n}{\partial p'^{(n)}_{j|i'}} \geq \frac{\partial J_n}{\partial p_{j'|i}^{(n)}}$$

Hence:

$$\frac{\partial J_n}{\partial p_{.|.}^{(n)}} \left(\partial E - \partial E' \right) \geq 0$$

Thus, the greedy algorithm finds the assignment that most increases J_n most under the children size constraints.

Moreover, $\frac{\partial J_n}{\partial p_{.|.}^{(n)}}$ is always positive for $L \leq M$ or $L \geq 2M(M-2)$. □

Proof of Lemma 2. Both J_n and J_T are defined as the sum of non-negative values which gives the lower-bound. We next derive the upper-bound on J_n. Recall:

$$J_n = \frac{2}{M} \sum_{j=1}^{M} \sum_{i=1}^{K} q_i^{(n)} |p_j^{(n)} - p_{j|i}^{(n)}| = \frac{2}{M} \sum_{j=1}^{M} \sum_{i=1}^{K} q_i^{(n)} \left| \sum_{l=1}^{K} q_l^{(n)} p_{j|l}^{(n)} - p_{j|i}^{(n)} \right|$$

since $p_j^{(n)} = \sum_{l=1}^{K} q_l^{(n)} p_{j|l}^{(n)}$. The objective J_n is maximized on the extremes of the $[0,1]$ interval. Thus, define the following two sets of indices:

$$O_j = \{i : i \in \{1,2,\ldots,K\}, p_{j|i}^{(n)} = 1\} \quad \text{and} \quad Z_j = \{i : i \in \{1,2,\ldots,K\}, p_{j|i}^{(n)} = 0\}.$$

We omit indexing these sets with n for the ease of notation. We continue as follows

$$\begin{aligned} J_n &\leq \frac{2}{M}\sum_{j=1}^{M}\left[\sum_{i\in O_j} q_i^{(n)}\left(1 - \sum_{l\in O_j} q_l^{(n)}\right) + \sum_{i\in Z_j} q_i^{(n)} \sum_{l\in O_j} q_l^{(n)}\right] \\ &= \frac{4}{M}\sum_{j=1}^{M}\left[\sum_{i\in O_j} q_i^{(n)} - \left(\sum_{i\in O_j} q_i^{(n)}\right)^2\right] \\ &= \frac{4}{M}\left[1 - \sum_{j=1}^{M}\left(\sum_{i\in O_j} q_i^{(n)}\right)^2\right], \end{aligned}$$

where the last inequality is the consequence of the following: $\sum_{j=1}^{M} p_j^{(n)} = 1$ and $p_j^{(n)} = \sum_{l=1}^{K} q_l^{(n)} p_{j|l}^{(n)} = \sum_{i\in O_j} q_i^{(n)}$, thus $\sum_{j=1}^{M}\sum_{i\in O_j} q_i^{(n)} = 1$. Apllying Jensen's ineqality to the last inequality obtained gives

$$\begin{aligned} J_n &\leq \frac{4}{M} - 4\left[\sum_{j=1}^{M}\left(\frac{1}{M}\sum_{i\in O_j} q_i^{(n)}\right)\right]^2 \\ &= \frac{4}{M}\left(1 - \frac{1}{M}\right) \end{aligned}$$

That ends the proof. ☐

Proof of Lemma 3. We start from proving that if the split in node n is perfectly balanced, i.e. $\forall_{j=\{1,2,\ldots,M\}} p_j^{(n)} = \frac{1}{M}$, and perfectly pure, i.e. $\forall_{\substack{j=\{1,2,\ldots,M\}\\ i=\{1,2,\ldots,K\}}} \min(p_{j|i}^{(n)}, 1 - p_{j|i}^{(n)}) = 0$, then J_n admits the highest value $J_n = \frac{4}{M}\left(1 - \frac{1}{M}\right)$. Since the split is maximally balanced we write:

$$J_n = \frac{2}{M}\sum_{j=1}^{M}\sum_{i=1}^{K} q_i^{(n)} \left|\frac{1}{M} - p_{j|i}^{(n)}\right|.$$

Since the split is maximally pure, each $p_{j|i}^{(n)}$ can only take value 0 or 1. As in the proof of previous lemma, define two sets of indices:

$$O_j = \{i : i \in \{1,2,\ldots,K\}, p_{j|i}^{(n)} = 1\} \quad \text{and} \quad Z_j = \{i : i \in \{1,2,\ldots,K\}, p_{j|i}^{(n)} = 0\}.$$

We omit indexing these sets with n for the ease of notation. Thus

$$\begin{aligned} J_n &= \frac{2}{M}\sum_{j=1}^{M}\left[\sum_{i\in O_j} q_i^{(n)}\left(1 - \frac{1}{M}\right) + \sum_{i\in Z_j} q_i^{(n)}\frac{1}{M}\right] \\ &= \frac{2}{M}\sum_{j=1}^{M}\left[\sum_{i\in O_j} q_i^{(n)}\left(1 - \frac{1}{M}\right) + \frac{1}{M}\left(1 - \sum_{i\in O_j} q_i^{(n)}\right)\right] \\ &= \frac{2}{M}\left(1 - \frac{2}{M}\right)\sum_{j=1}^{M}\sum_{i\in O_j} q_i^{(n)} + \frac{2}{M} \\ &= \frac{4}{M}\left(1 - \frac{1}{M}\right), \end{aligned}$$

where the last equality comes from the fact that $\sum_{j=1}^{M} p_j^{(n)} = 1$ and $p_j^{(n)} = \sum_{l=1}^{K} q_l^{(n)} p_{j|l}^{(n)} = \sum_{i\in O_j} q_i^{(n)}$, thus $\sum_{j=1}^{M}\sum_{i\in O_j} q_i^{(n)} = 1$.

Thus we are done with proving one induction direction. Next we prove that if J_n admits the highest value $J_n = \frac{4}{M}\left(1-\frac{1}{M}\right)$, then the split in node n is perfectly balanced, i.e. $\forall_{j=\{1,2,\dots,M\}} p_j^{(n)} = \frac{1}{M}$, and perfectly pure, i.e. $\forall_{\substack{j=\{1,2,\dots,M\}\\ i=\{1,2,\dots,K\}}} \min(p_{j|i}^{(n)}, 1-p_{j|i}^{(n)}) = 0$.

Without loss of generality assume each $q_i^{(n)} \in (0,1)$. The objective J_n is certainly maximized in the extremes of the interval $[0,1]$, where each $p_{j|i}^{(n)}$ is either 0 or 1. Also, at maximum it cannot be that for any given j, all $p_{j|i}^{(n)}$'s are 0 or all $p_{j|i}^{(n)}$'s are 1. The function $J(h)$ is differentiable in these extremes. Next, define three sets of indices:

$$\mathcal{A}_j = \{i : \sum_{l=1}^{K} q_i^{(n)} p_{j|l}^{(n)} \geq p_{j|i}^{(n)}\} \quad \text{and} \quad \mathcal{B}_j = \{i : \sum_{l=1}^{K} q_i^{(n)} p_{j|l}^{(n)} < p_{j|i}^{(n)}\} \quad \text{and} \quad \mathcal{C}_j = \{i : \sum_{l=1}^{K} q_i^{(n)} p_{j|l}^{(n)} > p_{j|i}^{(n)}\}.$$

We omit indexing these sets with n for the ease of notation. Objective J_n can then be re-written as

$$J_n = \frac{2}{M}\sum_{j=1}^{M}\left[\sum_{i\in\mathcal{A}_j} q_i^{(n)}\left(\sum_{l=1}^{K} q_i^{(n)} p_{j|l}^{(n)} - p_{j|i}^{(n)}\right) + 2\sum_{i\in\mathcal{B}_j} q_i^{(n)}\left(p_{j|i}^{(n)} - \sum_{l=1}^{K} q_i^{(n)} p_{j|l}^{(n)}\right)\right],$$

We next compute the derivatives of J_n with respect to $p_{j|z}^{(n)}$, where $z = \{1,2,\dots,K\}$, everywhere where the function is differentiable and obtain

$$\frac{\partial J_n}{\partial p_{j|z}^{(n)}} = \begin{cases} 2q_z^{(n)}(\sum_{i\in\mathcal{C}_j} q_i^{(n)} - 1) & \text{if } z \in \mathcal{C}_j \\ 2q_z^{(n)}(1 - \sum_{i\in\mathcal{B}_j} q_i^{(n)}) & \text{if } z \in \mathcal{B}_j \end{cases},$$

Note that in the extremes of the interval $[0,1]$ where J_n is maximized, it cannot be that $\sum_{i\in\mathcal{C}_j} q_i^{(n)} = 1$ or $\sum_{i\in\mathcal{B}_j} q_i^{(n)} = 1$ thus the gradient is non-zero. This fact and the fact that J_n is convex imply that J_n can *only* be maximized at the extremes of the $[0,1]$ interval. Thus if J_n admits the highest value, then the node split is perfectly pure. We still need to show that if J_n admits the highest value, then the node split is also perfectly balanced. We give a proof by contradiction, thus we assume that at least for one value of j, $p_j^{(n)} \neq \frac{1}{M}$, or in other words if we decompose each $p_j^{(n)}$ as $p_j^{(n)} = \frac{1}{M} + x_j$, then at least for one value of j, $x_j \neq 0$. Lets once again define two sets of indices (we omit indexing x_j and these sets with n for the ease of notation):

$$O_j = \{i : i \in \{1,2,\dots,K\}, p_{j|i}^{(n)} = 1\} \quad \text{and} \quad Z_j = \{i : i \in \{1,2,\dots,K\}, p_{j|i}^{(n)} = 0\},$$

and recall that $p_j^{(n)} = \sum_{l=1}^{K} q_l^{(n)} p_{j|l}^{(n)} = \sum_{i \in O_j} q_i^{(n)}$. We proceed as follows

$$\begin{aligned}
\frac{4}{M}\left(1 - \frac{1}{M}\right) = J_n &= \frac{2}{M}\sum_{j=1}^{M}\left[\sum_{i \in O_j} q_i^{(n)}(1 - p_j^{(n)}) + \sum_{i \in Z_j} q_i^{(n)} p_j^{(n)}\right] \\
&= \frac{2}{M}\sum_{j=1}^{M}\left[p_j^{(n)}(1 - p_j^{(n)}) + p_j^{(n)}(1 - p_j^{(n)})\right] \\
&= \frac{4}{M}\sum_{j=1}^{M}\left[p_j^{(n)} - (p_j^{(n)})^2\right] \\
&= \frac{4}{M}\left[1 - \sum_{j=1}^{M}(p_j^{(n)})^2\right] \\
&= \frac{4}{M}\left[1 - \sum_{j=1}^{M}\left(\frac{1}{M} + x_j\right)^2\right] \\
&= \frac{4}{M}\left(1 - \frac{1}{M} - \frac{2}{M}\sum_{j=1}^{M} x_j - \sum_{j=1}^{M} x_j^2\right) \\
&< \frac{4}{M}\left(1 - \frac{1}{M}\right)
\end{aligned}$$

Thus we obtain the contradiction which ends the proof. □

Proof of Lemma 4. Since we node that the split is perfectly pure, then each $p_{j|i}^{(n)}$ is either 0 or 1. Thus we define two sets

$$O_j = \{i : i \in \{1, 2, \ldots, K\}, p_{j|i}^{(n)} = 1\} \quad \text{and} \quad Z_j = \{i : i \in \{1, 2, \ldots, K\}, p_{j|i}^{(n)} = 0\}.$$

and thus

$$J_n = \frac{2}{M}\sum_{j=1}^{M}\left[\sum_{i \in O_j} q_i^{(n)}\,(1 - p_j) + \sum_{i \in Z_j} q_i^{(n)} p_j\right]$$

Note that $p_j = \sum_{i \in O_j} q_i^{(n)}$. Then

$$J_n = \frac{2}{M}\sum_{j=1}^{M}\left[p_j\,(1 - p_j) + (1 - p_j)p_j\right] = \frac{4}{M}\sum_{j=1}^{M} p_j\,(1 - p_j) = \frac{4}{M}\left(1 - \sum_{j=1}^{M} p_j^2\right)$$

and thus

$$\sum_{j=1}^{M} p_j^2 = 1 - \frac{M J_n}{4}. \tag{16}$$

Lets express p_j as $p_j = \frac{1}{M} + \epsilon_j$, where $\epsilon_j \in [-\frac{1}{M}, 1 - \frac{1}{M}]$. Then

$$\sum_{j=1}^{M} p_j^2 = \sum_{j=1}^{M}\left(\frac{1}{M} + \epsilon_j\right)^2 = \frac{1}{M} + \frac{2}{M}\sum_{j=1}^{M}\epsilon_j + \sum_{j=1}^{M}\epsilon_j^2 = \frac{1}{M} + \sum_{j=1}^{M}\epsilon_j^2, \tag{17}$$

since $\frac{2}{M}\sum_{j=1}^{M}\epsilon_j = 0$. Thus combining Equation 16 and 17

$$\frac{1}{M} + \sum_{j=1}^{M}\epsilon_j^2 = 1 - \frac{M J_n}{4}$$

and thus

$$\sum_{j=1}^{M} \epsilon_j^2 = 1 - \frac{1}{M} - \frac{MJ_n}{4}.$$

The last statement implies that

$$\max_{j=1,2,\ldots,M} \epsilon_j \leq \sqrt{1 - \frac{1}{M} - \frac{MJ_n}{4}},$$

which is equivalent to

$$\min_{j=1,2,\ldots,M} p_j = \frac{1}{M} - \max_j \epsilon_j \geq \frac{1}{M} - \sqrt{1 - \frac{1}{M} - \frac{MJ_n}{4}} = \frac{1}{M} - \frac{\sqrt{M(J^* - J_n)}}{2}.$$

□

Proof of Lemma 5. Since the split is perfectly balanced we have the following:

$$J_n = \frac{2}{M} \sum_{j=1}^{M} \sum_{i=1}^{K} q_i^{(n)} \left| \frac{1}{M} - p_{j|i}^{(n)} \right| = \frac{2}{M} \sum_{i=1}^{K} \sum_{j=1}^{M} q_i^{(n)} \left| \frac{1}{M} - p_{j|i}^{(n)} \right|$$

Define two sets

$$\mathcal{A}_i = \{j : j \in \{1, 2, \ldots, K\}, p_{j|i}^{(n)} < \frac{1}{M}\} \quad \text{and} \quad \mathcal{B}_i = \{j : j \in \{1, 2, \ldots, K\}, p_{j|i}^{(n)} \geq \frac{1}{M}\}.$$

Then

$$\begin{aligned}
J_n &= \frac{2}{M} \sum_{i=1}^{K} \left[\sum_{j \in \mathcal{A}_i} q_i^{(n)} \left(\frac{1}{M} - p_{j|i}^{(n)} \right) + \sum_{j \in \mathcal{B}_i} q_i^{(n)} \left(p_{j|i}^{(n)} - \frac{1}{M} \right) \right] \\
&= \frac{2}{M} \sum_{i=1}^{K} q_i^{(n)} \left[\sum_{j \in \mathcal{A}_i} \left(\frac{1}{M} - p_{j|i}^{(n)} \right) + \sum_{j \in \mathcal{B}_i} \left(p_{j|i}^{(n)} - \frac{1}{M} \right) \right] \\
&= \frac{2}{M} \sum_{i=1}^{K} q_i^{(n)} \left[\sum_{j \in \mathcal{A}_i} \left(\frac{1}{M} - p_{j|i}^{(n)} \right) + \sum_{j \in \mathcal{B}_i} \left((1 - \frac{1}{M}) - (1 - p_{j|i}^{(n)}) \right) \right]
\end{aligned}$$

Recall that the optimal value of J_n is:

$$J^* = \frac{4}{M}\left(1 - \frac{1}{M}\right) = \frac{2}{M} \sum_{i=1}^{N} q_i^{(n)} \left[(M-1)\frac{1}{M} + \left(1 - \frac{1}{M}\right) \right] = \frac{2}{M} \sum_{i=1}^{N} q_i^{(n)} \left[\left(\sum_{j \in \mathcal{A}_i \cup \mathcal{B}_i} \frac{1}{M} \right) - \frac{1}{M} + \left(1 - \frac{1}{M}\right) \right]$$

Note $\mathcal{A}_i$ can have at most $M-1$ elements. Furthermore, $\forall j \in \mathcal{A}_i, p_{j|i}^{(n)} < 1 - p_{j|i}^{(n)}$. Then, we have:

$$J^* - J^n = \frac{2}{M} \sum_{i=1}^{K} q_i^{(n)} \left[\sum_{j \in \mathcal{A}_i} p_{j|i}^{(n)} + \sum_{j \in \mathcal{B}_i} \left((1 - p_{j|i}^{(n)}) + \frac{1}{M} - (1 - \frac{1}{M}) \right) - \frac{1}{M} + \left(1 - \frac{1}{M}\right) \right]$$

Hence, since $\mathcal{B}_i$ has at least one element:

$$\begin{aligned}
J^* - J^n &\geq \frac{2}{M} \sum_{i=1}^{K} q_i^{(n)} \left[\sum_{j \in \mathcal{A}_i} p_{j|i}^{(n)} + \sum_{j \in \mathcal{B}_i} \left(1 - p_{j|i}^{(n)} \right) \right] \\
&\geq \frac{2}{M} \sum_{i=1}^{K} q_i^{(n)} \left[\sum_{j=1}^{M} \min(p_{j|i}^{(n)}, 1 - p_{j|i}^{(n)}) \right] \\
&\geq 2\alpha
\end{aligned}$$

□

Proof of Theorem 1. Let the weight of the tree leaf be defined as the probability that a randomly chosen data point x drawn from some fixed target distribution $\mathcal{P}$ reaches this leaf. Suppose at time step t, n is the heaviest leaf and has weight w. Consider splitting this leaf to M children $n_1, n_2, \ldots, n_M$. Let the weight of the j^{th} child be denoted as w_j. Also for the ease of notation let p_j refer to $p_j^{(n)}$ (recall that $\sum_{j=1}^{m} p_j = 1$) and $p_{j|i}$ refer to $p_{j|i}^{(n)}$, and furthermore let q_i be the shorthand for $q_i^{(n)}$. Recall that $p_j = \sum_{i=1}^{K} q_i p_{j|i}$ and $\sum_{i=1}^{K} q_i = 1$. Notice that for any $j = \{1, 2, \ldots, M\}$, $w_j = wp_j$. Let $\boldsymbol{q}$ be the k-element vector with i^{th} entry equal to q_i. Define the following function: $\tilde{G}^e(\boldsymbol{q}) = \sum_{i=1}^{K} q_i \ln\left(\frac{1}{q_i}\right)$. Recall the expression for the entropy of tree leaves: $G^e = \sum_{l \in \mathcal{L}} w_l \sum_{i=1}^{K} q_i^{(l)} \ln\left(\frac{1}{q_i^{(l)}}\right)$, where $\mathcal{L}$ is a set of all tree leaves. Before the split the contribution of node n to G^e was equal to $w\tilde{G}^e(\boldsymbol{q})$. Note that for any $j = \{1, 2, \ldots, M\}$, $q_i^{(n_j)} = \frac{q_i p_{j|i}}{p_j}$ is the probability that a randomly chosen x drawn from $\mathcal{P}$ has label i given that x reaches node n_j. For brevity, let $q_i^{n^j}$ be denoted as $q_{j,i}$. Let $\boldsymbol{q}_j$ be the k-element vector with i^{th} entry equal to $q_{j,i}$. Notice that $\boldsymbol{q} = \sum_{j=1}^{M} p_j \boldsymbol{q}_j$. After the split the contribution of the same, now internal, node n changes to $w \sum_{j=1}^{M} p_j \tilde{G}^e(\boldsymbol{q}_j)$. We denote the difference between the contribution of node n to the value of the entropy-based objectives in times t and $t+1$ as

$$\Delta_t^e := G_t^e - G_{t+1}^e = w\left[\tilde{G}^e(\boldsymbol{q}) - \sum_{j=1}^{M} p_j \tilde{G}^e(\boldsymbol{q}_j)\right]. \tag{18}$$

The entropy function $\tilde{G}^e$ is strongly concave with respect to l_1-norm with modulus 1, thus we extend the inequality given by Equation 7 in (Choromanska et al., 2016) by applying Theorem 5.2. from (Azocar et al., 2011) and obtain the following bound

$$\begin{aligned}
\Delta_t^e &= w\left[\tilde{G}^e(\boldsymbol{q}) - \sum_{j=1}^{M} p_j \tilde{G}^e(\boldsymbol{q}_j)\right] \\
&\geq w\frac{1}{2}\sum_{j=1}^{M} p_j \|q_j - \sum_{l=1}^{M} p_l q_l\|_1^2 \\
&= w\frac{1}{2}\sum_{j=1}^{M} p_j \left(\sum_{i=1}^{K}\left|\frac{q_i p_{j|i}}{p_j} - \sum_{l=1}^{M} p_l \frac{q_i p_{l|i}}{p_l}\right|\right)^2 \\
&= w\frac{1}{2}\sum_{j=1}^{M} p_j \left(\sum_{i=1}^{K} q_i\left|\frac{p_{j|i}}{p_j} - \sum_{l=1}^{M} p_{l|i}\right|\right)^2 \\
&= w\frac{1}{2}\sum_{j=1}^{M} p_j \left(\sum_{i=1}^{K} q_i\left|\frac{p_{j|i}}{p_j} - 1\right|\right)^2 \\
&= w\frac{1}{2}\sum_{j=1}^{M} \frac{1}{p_j} \left(\sum_{i=1}^{K} q_i\left|p_{j|i} - p_j\right|\right)^2.
\end{aligned}$$

Before proceeding, we will bound each p_j. Note that by the *Weak Hypothesis Assumption* we have

$$\gamma \in \left[\frac{M}{2}\min_{j=1,2,\ldots,M} p_j, 1 - \frac{M}{2}\min_{j=1,2,\ldots,M} p_j\right],$$

thus

$$\min_{j=1,2,\ldots,M} p_j \geq \frac{2\gamma}{M},$$

thus all p_js are such that $p_j \geq \frac{2\gamma}{M}$. Thus

$$\max_{j=1,2,\ldots,M} p_j \leq 1 - \frac{2\gamma}{M}(M-1) = \frac{M(1-2\gamma)+2\gamma}{M}.$$

Thus all p_js are such that $p_j \leq \frac{M(1-2\gamma)+2\gamma}{M}$.

$$\begin{aligned}
\Delta_t^e &\geq w\frac{M^2}{2[(M(1-2\gamma)+2\gamma]}\sum_{j=1}^{M}\frac{1}{M}\left(\sum_{i=1}^{K} q_i \left|p_{j|i}-p_j\right|\right)^2 \\
&\geq w\frac{M^2}{2[(M(1-2\gamma)+2\gamma]}\left(\sum_{j=1}^{M}\frac{1}{M}\sum_{i=1}^{K} q_i \left|p_{j|i}-p_j\right|\right)^2 \\
&= w\frac{M^2}{8[(M(1-2\gamma)+2\gamma]}\left(\frac{2}{M}\sum_{j=1}^{M}\sum_{i=1}^{K} q_i \left|p_{j|i}-p_j\right|\right)^2 \\
&= \frac{M^2}{[(M(1-2\gamma)+2\gamma]}\frac{wJ_n^2}{8},
\end{aligned}$$

where the last inequality is a consequence of Jensen's inequality. w can further be lower-bounded by noticing the following

$$G_t^e = \sum_{l\in\mathcal{L}} w_l \sum_{i=1}^{K} q_i^{(l)} \ln\left(\frac{1}{q_i^{(l)}}\right) \leq \sum_{l\in\mathcal{L}} w_l \ln K \leq w \ln K \sum_{l\in\mathcal{L}} 1 = [t(M-1)+1]w\ln K \leq (t+1)(M-1)w\ln K,$$

where the first inequality results from the fact that uniform distribution maximizes the entropy.

This gives the lower-bound on Δ_t^e of the following form:

$$\Delta_t^e \geq \frac{M^2 G_t^e J_n^2}{8(t+1)[M(1-2\gamma)+2\gamma](M-1)\ln K},$$

and by using *Weak Hypothesis Assumption* we get

$$\Delta_t^e \geq\geq \frac{M^2 G_t^e \gamma^2}{8(t+1)[M(1-2\gamma)+2\gamma](M-1)\ln K}$$

Following the recursion of the proof in Section 3.2 in (Choromanska et al., 2016) (note that in our case $G_1^e \leq 2(M-1)\ln K$), we obtain that under the *Weak Hypothesis Assumption*, for any $\kappa \in [0, 2(M-1)\ln K]$, to obtain $G_t^e \leq \kappa$ it suffices to make

$$t \geq \left(\frac{2(M-1)\ln K}{\kappa}\right)^{\frac{16[M(1-2\gamma)+2\gamma](M-1)\ln K}{M^2\log_2 e\gamma^2}}$$

splits. We next proceed to directly proving the error bound. Denote $w(l)$ to be the probability that a data point x reached leaf l. Recall that $q_i^{(l)}$ is the probability that the data point x corresponds to label i given that x reached l, i.e. $q_i^{(l)} = P(y(x)=i|x \text{ reached } l)$. Let the label assigned to the leaf be the majority label and thus lets assume that the leaf is assigned to label i if and only if the following is true $\forall_{\substack{z=\{1,2,\dots,k\}\\ z\neq i}} q_i^{(l)} \geq q_z^{(l)}$. Therefore we can write that

$$\begin{aligned}
\epsilon(\mathcal{T}) &= \sum_{i=1}^{K} P(t(x)=i, y(x)\neq i) &(19)\\
&= \sum_{l\in\mathcal{L}} w(l)\sum_{i=1}^{K} P(t(x)=i, y(x)\neq i|x \text{ reached } l) \\
&= \sum_{l\in\mathcal{L}} w(l)\sum_{i=1}^{K} P(y(x)\neq i|t(x)=i, x \text{ reached } l)P(t(x)=i|x \text{ reached } l) \\
&= \sum_{l\in\mathcal{L}} w(l)(1-\max(q_1^{(l)}, q_2^{(l)}, \dots, q_K^{(l)}))\sum_{i=1}^{K} P(t(x)=i|x \text{ reached } l) \\
&= \sum_{l\in\mathcal{L}} w(l)(1-\max(q_1^{(l)}, q_2^{(l)}, \dots, q_K^{(l)})) &(20)
\end{aligned}$$

Consider again the Shannon entropy $G(\mathcal{T})$ of the leaves of tree $\mathcal{T}$ that is defined as

$$G^e(\mathcal{T}) = \sum_{l\in\mathcal{L}} w(l) \sum_{i=1}^{K} q_i^{(l)} \log_2 \frac{1}{q_i^{(l)}}. \tag{21}$$

Let $i_l = \arg\max_{i=\{1,2,\ldots,K\}} q_i^{(l)}$. Note that

$$\begin{aligned}
G^e(\mathcal{T}) &= \sum_{l\in\mathcal{L}} w(l) \sum_{i=1}^{K} q_i^{(l)} \log_2 \frac{1}{q_i^{(l)}} \\
&\geq \sum_{l\in\mathcal{L}} w(l) \sum_{\substack{i=1\\ i\neq i_l}}^{K} q_i^{(l)} \log_2 \frac{1}{q_i^{(l)}} \\
&\geq \sum_{l\in\mathcal{L}} w(l) \sum_{\substack{i=1\\ i\neq i_l}}^{K} q_i^{(l)} \\
&= \sum_{l\in\mathcal{L}} w(l)(1 - \max(q_1^{(l)}, q_2^{(l)}, \ldots, q_K^{(l)})) \\
&= \epsilon(\mathcal{T}),
\end{aligned} \tag{22}$$

where the last inequality comes from the fact that $\forall_{\substack{i=\{1,2,\ldots,K\}\\ i\neq i_l}} q_i^{(l)} \leq 0.5$ and thus $\forall_{\substack{i=\{1,2,\ldots,K\}\\ i\neq i_l}} \frac{1}{q_i^{(l)}} \in [2; +\infty]$ and consequently $\forall_{\substack{i=\{1,2,\ldots,K\}\\ i\neq i_l}} \log_2 \frac{1}{q_i^{(l)}} \in [1; +\infty]$.

We next use the proof of Theorem 6 in (Choromanska et al., 2016). The proof modifies only slightly for our purposes and thus we only list these modifications below.

- Since we define the Shannon entropy through logarithm with base 2 instead of the natural logarithm, the right hand side of inequality (2.6) in (Shalev-Shwartz, 2012) should have an additional multiplicative factor equal to $\frac{1}{\ln 2}$ and thus the right-hand side of the inequality stated in Lemma 14 has to have the same multiplicative factor.
- For the same reason as above, the right-hand side of the inequality in Lemma 9 should take logarithm with base 2 of k instead of the natural logarithm of k.

Propagating these changes in the proof of Theorem 6 results in the statement of Theorem 1.

$\square$

Proof of Corollary 1. Note that the lower-bound on Δ_t^e from the previous prove could be made tighter as follows:

$$\begin{aligned}
\Delta_t^e &\geq w\frac{1}{2}\sum_{j=1}^{M}\frac{1}{p_j}\left(\sum_{i=1}^{K} q_i \left|p_{j|i} - p_j\right|\right)^2 \\
&= w\frac{M^2}{2}\sum_{j=1}^{M}\frac{1}{M}\left(\sum_{i=1}^{K} q_i \left|p_{j|i} - p_j\right|\right)^2 \\
&\geq w\frac{M^2}{2}\left(\sum_{j=1}^{M}\frac{1}{M}\sum_{i=1}^{K} q_i \left|p_{j|i} - p_j\right|\right)^2 \\
&= w\frac{M^2}{8}\left(\frac{2}{M}\sum_{j=1}^{M}\sum_{i=1}^{K} q_i \left|p_{j|i} - p_j\right|\right)^2 \\
&= \frac{M^2 w J_n^2}{8},
\end{aligned}$$

d	Model	Arity	Prec	Rec	Train	Test
50	TagSpace	-	30.1	-	3h8	6h
	FastText	2	27.2	4.17	**8m**	**1m**
	Huffman Tree	5	28.3	4.33	**8m**	**1m**
		20	29.9	4.58	10m	3m
	Learned Tree	5	31.6	4.85	18m	**1m**
		20	**32.1**	**4.92**	30m	3m
200	TagSpace	-	35.6	-	5h32	15h
	FastText	2	35.2	5.4	**12m**	**1m**
	Huffman Tree	5	35.8	5.5	13m	2m
		20	36.4	5.59	18m	3m
	Learned Tree	5	36.1	5.53	35m	3m
		20	**36.6**	**5.61**	45m	8m

Table 3. Classification performance on the YFCC100M dataset.

Model	perp.	train ms/batch	test ms/batch
Random Tree	172	*5.1*	*2.7*
Flat soft-max	151	*11.5*	*5.1*
Learned Tree	159	*6.3*	*2.6*

Table 4. Comparison of a flat soft-max to a 25-ary hierarchical soft-max (learned, random and heuristic-based tree).

where the first inequality was taken from the proof of Theorem 1 and the following equality follows from the fact that each node is balanced. By next following exactly the same steps as shown in the proof of Theorem 1 we obtain the corollary. □

11. Experimental Setting

11.1. Classification

For the YFCC100M experiments, we learned our models with SGD with a linearly decreasing rate for five epochs. We run a hyper-parameter search on the learning rate (in $\{0.01, 0.02, 0.05, 0.1, 0.25, 0.5\}$). In the learned tree settings, the learning rate stays constant for the first half of training, during which the AssignLabels() routine is called 50 times. We run the experiments in a Hogwild data-parallel setting using 12 threads on an Intel Xeon E5-2690v4 2.6GHz CPU. At prediction time, we perform a truncated depth first search to find the most likely label (using the same idea as in a branch-and-bound algorithm: if a node score is less than that of the best current label, then all of its descendants are out).

11.2. Density Estimation

In our experiments, we use a context window size of 4. We optimize the objectives with Adagrad, run a hyper-parameter search on the batch size (in $\{32, 64, 128\}$) and learning rate (in $\{0.01, 0.02, 0.05, 0.1, 0.25, 0.5\}$). The hidden representation dimension is 200. In the learned tree settings, the AssignLabels() routine is called 50 times per epoch. We used a 12GB NVIDIA GeForce GTX TITAN GPU and all tree-based models are 65-ary for the Gutenberg data and 25-ary for Pen TreeBank. Table 4 provides the perplexity and speed results on the PTB text.

For the Cluster Tree, we learn dimension 50 word embeddings with FastTree for 5 epochs using a hierarchical softmax loss, then obtain $45 = 65^2$ centroids using the ScikitLearn implementation of MiniBatchKmeans, and greedily assign words to clusters until full (when a cluster has 65 words).

Algorithm 3 Label Assignment Algorithm under Depth Constraint

Input Node statistics, max depth D
  Paths from root to labels: $\mathcal{P} = (\mathbf{c}^i)_{i=1}^K$
  node ID n and depth d
  List of labels currently reaching the node
Ouput Updated paths
  Lists of labels now assigned to each of n's children under depth constraints

procedure AssignLabels (labels, n, d)
  // first, compute $p_j^{(n)}$ and $p_{j|i}^{(n)}$. $\odot$ is the element-wise
  // multiplication
  $\mathbf{p}_0^{avg} \leftarrow \mathbf{0}$
  count $\leftarrow 0$
  for i in labels **do**
    $\mathbf{p}_0^{avg} \leftarrow \mathbf{p}_0^{avg} + \text{SumProbas}_{n,i}$
    count $\leftarrow$ count + $\text{Counts}_{n,i}$
    $\mathbf{p}_i^{avg} \leftarrow \text{SumProbas}_{n,i}/\text{Counts}_{n,i}$
  $\mathbf{p}_0^{avg} \leftarrow \mathbf{p}_0^{avg}/\text{count}$
  // then, assign each label to a child of n under depth
  // constraints
  unassigned $\leftarrow$ labels
  full $\leftarrow \emptyset$
  for $j = 1$ to M **do**
    $\text{assigned}_j \leftarrow \emptyset$
  while unassigned $\neq \emptyset$ **do**
    // $\frac{\partial J_n}{\partial p_{j|i}^{(n)}}$ is given in Equation 10
    $(i^*, j^*) \leftarrow \underset{i \in \text{unassigned}, j \notin \text{full}}{\text{argmax}} \left(\frac{\partial J_n}{\partial p_{j|i}^{(n)}} \right)$
    $\mathbf{c}_d^{i^*} \leftarrow (n, j^*)$
    $\text{assigned}_{j^*} \leftarrow \text{assigned}_{j^*} \cup \{i^*\}$
    unassigned $\leftarrow$ unassigned $\setminus \{i^*\}$
    if $|\text{assigned}_{j^*}| = M^{D-d}$ **then**
      full $\leftarrow$ full $\cup \{j^*\}$
  for $j = 1$ to M **do**
    AssignLabels (assigned_j, $\text{child}_{n,j}$, $d+1$)
  return assigned

Leaf 229	Leaf 230	Leaf 300	Leaf 231
suggested	vegas	payments	operates
watched	&	buy-outs	includes
created	calif.	swings	intends
violated	park	gains	makes
introduced	n.j.	taxes	means
discovered	conn.	operations	helps
carried	pa.	profits	seeks
described	pa.	penalties	reduces
accepted	ii	relations	continues
listed	d.	liabilities	fails
…	…	…	…

Table 5. Example of labels reaching leaf nodes in the final tree. We can identify a leaf for 3rd person verbs, one for past participates, one for plural nouns, and one (loosely) for places.