Supplementary Material

A. Proofs

In the theorectical analysis, we fix $s_K(\boldsymbol{x}, \boldsymbol{\theta}) = 0$. Then, we only need to consider $\mathcal{C}_{\boldsymbol{x}} \cup \mathcal{N}_{\boldsymbol{x}} = \{1, \cdots, K-1\}$. Now, the normalization factor becomes

$$E(\boldsymbol{x}, j) = 1 + \sum_{k' \in \mathcal{C}_{\boldsymbol{x}}} e^{s_{k'}(\boldsymbol{x}, \boldsymbol{\theta})} + e^{s_j(\boldsymbol{x}, \boldsymbol{\theta})}/q_{\boldsymbol{x}}(j),$$

with some sampled class $j \in \mathcal{N}_{\boldsymbol{x}}$. Now, we can rewrite R and $\hat{R}$ as

$$R(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}} \sum_{k \in \mathcal{C}_{\boldsymbol{x}}} p(y=k|\boldsymbol{x}) \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_{\boldsymbol{x}}(j) \log \frac{e^{s_k(\boldsymbol{x}, \boldsymbol{\theta})}}{E(\boldsymbol{x}, j)} + \sum_{k \in \mathcal{N}_{\boldsymbol{x}}} p(y=k|\boldsymbol{x}) \log \frac{e^{s_k(\boldsymbol{x}, \boldsymbol{\theta})}}{E(\boldsymbol{x}, k)} + p(y=K|\boldsymbol{x}) \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_{\boldsymbol{x}}(j) \log \frac{1}{E(\boldsymbol{x}, j)}.$$

$$\hat{R}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{k \in \mathcal{C}_{\boldsymbol{x}_i}} \mathbb{I}(y_i = k) \sum_{j \in \mathcal{C}_{\boldsymbol{x}_i}} q_{\boldsymbol{x}_i}(j) \log \frac{e^{s_k(\boldsymbol{x}_i, \boldsymbol{\theta})}}{E(\boldsymbol{x}_i, j)} + \sum_{k \in \mathcal{N}_{\boldsymbol{x}_i}} \mathbb{I}(y_i = k) \log \frac{e^{s_k(\boldsymbol{x}_i, \boldsymbol{\theta})}}{E(\boldsymbol{x}_i, k)} + \mathbb{I}(y_i = K) \sum_{j \in \mathcal{C}_{\boldsymbol{x}_i}} q_{\boldsymbol{x}_i}(j) \log \frac{1}{E(\boldsymbol{x}_i, j)} \right].$$

In the proofs, we will use point-wise notations p_k, s_k, q_k and E_k to represent $p(y=k|\boldsymbol{x})$, $s_k(\boldsymbol{x}, \boldsymbol{\theta})$, $q_{\boldsymbol{x}}(k)$ and $E(\boldsymbol{x}, k)$ for simplicity.

A.1. Useful Lemma

We will need the following lemma in our analysis.

Lemma 1. *For any norm $\|\cdot\|$ defined on the parameter space of $\boldsymbol{\theta}$, assume the quantities $\|\nabla_{\boldsymbol{\theta}} s_k\|$, $\|\nabla^2_{\boldsymbol{\theta}} s_k\|$ and $\|\nabla^3_{\boldsymbol{\theta}} s_k\|$ for $k = 1, \cdots, K-1$ are bounded. Then, for any compact set $\mathbb{S}$ defined on the parameter space, we have*

$$\sup_{\boldsymbol{\theta} \in \mathbb{S}} |\hat{R}_n(\boldsymbol{\theta}) - R(\boldsymbol{\theta})| \xrightarrow{p} 0, \quad \sup_{\boldsymbol{\theta} \in \mathbb{S}} \|\nabla \hat{R}_n(\boldsymbol{\theta}) - \nabla R(\boldsymbol{\theta})\| \xrightarrow{p} 0, \quad \textit{and} \quad \sup_{\boldsymbol{\theta} \in \mathbb{S}} \|\nabla^2 \hat{R}_n(\boldsymbol{\theta}) - \nabla^2 R(\boldsymbol{\theta})\| \xrightarrow{p} 0.$$

Proof. For fixed $\boldsymbol{\theta}$, let

$$\psi(\boldsymbol{x}, y, \boldsymbol{\theta}) = \sum_{k \in \mathcal{C}_{\boldsymbol{x}}} \mathbb{I}(y=k) \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_j \log \frac{e^{s_k}}{1 + \sum_{k' \in C_{\boldsymbol{x}_i}} e^{s_{k'}} + \frac{e^{s_j}}{q_j}} + \mathbb{I}(y=K) \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_j \log \frac{1}{1 + \sum_{k' \in C_{\boldsymbol{x}}} e^{s_{k'}} + \frac{e^{s_j}}{q_j}}$$
$$+ \sum_{k \in \mathcal{N}_{\boldsymbol{x}}} \mathbb{I}(y=k) \log \frac{e^{s_k}}{1 + \sum_{k' \in C_{\boldsymbol{x}}} e^{s_{k'}} + \frac{e^{s_k}}{q_k}}.$$

Then we have $\hat{R}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \psi(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\theta})$ and $R(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}, y}\ \psi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})$. By the Law of Large Numbers, we know that $\hat{R}_n(\boldsymbol{\theta})$ converges point-wisely to $R(\boldsymbol{\theta})$ in probability.

According to the assumption, there exists a constant $M > 0$ such that

$$\|\nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{x}, y, \boldsymbol{\theta})\| \leq \sum_{k=1}^{K-1} \|\nabla_{\boldsymbol{\theta}} s_k\| \leq M.$$

Given any $\epsilon > 0$, we may find a finite cover $\mathbb{S}_\epsilon \subset \mathbb{S}$ so that for any $\boldsymbol{\theta} \in \mathbb{S}$, there exists $\boldsymbol{\theta}' \in \mathbb{S}_\epsilon$ such that $|\psi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) - \psi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}')| \leq M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \epsilon$. Since $\mathbb{S}_\epsilon$ is finite, as $n \to \infty$, $\sup_{\boldsymbol{\theta} \in \mathbb{S}_\epsilon} |\hat{R}_n(\boldsymbol{\theta}) - R(\boldsymbol{\theta})|$ converges to 0 in probability. Therefore, as $n \to \infty$, with probability 1, we have

$$\sup_{\boldsymbol{\theta} \in \mathbb{S}} |\hat{R}_n(\boldsymbol{\theta}) - R(\boldsymbol{\theta})| < 2\epsilon + \sup_{\boldsymbol{\theta} \in \mathbb{S}_\epsilon} |\hat{R}_n(\boldsymbol{\theta}) - R(\boldsymbol{\theta})| \to 2\epsilon.$$

Let $\epsilon \to 0$, we obtain the first bound. The second and the third bounds can be similarly obtained. □

A.2. Proof of Theorem 1

Proof. R can be re-written as

$$R = \mathbb{E}_{\boldsymbol{x}} \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_j \left(\sum_{k \in \mathcal{C}_{\boldsymbol{x}}} p_k \log \frac{e^{s_k}}{1 + \sum_{k' \in \mathcal{C}_{\boldsymbol{x}}} e^{s_{k'}} + e^{s_j}/q_j} + p_K \log \frac{1}{1 + \sum_{k' \in \mathcal{C}_{\boldsymbol{x}}} e^{s_{k'}} + e^{s_j}/q_j} + \frac{p_j}{q_j} \log \frac{e^{s_j}}{1 + \sum_{k' \in \mathcal{C}_{\boldsymbol{x}}} e^{s_{k'}} + e^{s_j}/q_j} \right).$$

For $i \in \mathcal{C}_{\boldsymbol{x}}$, we have

$$\begin{aligned}
\nabla_{s_i} R &= \mathbb{E}_{\boldsymbol{x}} \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_j \Bigg[p_i \left(1 - \frac{e^{s_i}}{1 + \sum_{k' \in \mathcal{C}_x} e^{s_{k'}} + e^{s_j}/q_j}\right) - \sum_{k \neq i \in \mathcal{C}_{\boldsymbol{x}}} p_k \frac{e^{s_i}}{1 + \sum_{k' \in \mathcal{C}_x} e^{s_{k'}} + e^{s_j}/q_j} \\
&\quad - p_K \frac{e^{s_i}}{1 + \sum_{k' \in \mathcal{C}_x} e^{s_{k'}} + e^{s_j}/q_j} - p_j/q_j \frac{e^{s_i}}{1 + \sum_{k' \in \mathcal{C}_x} e^{s_{k'}} + e^{s_j}/q_j} \Bigg] \\
&= \mathbb{E}_{\boldsymbol{x}} \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_j \left[p_i - \left(p_K + \sum_{k \in \mathcal{C}_x} p_k + p_j/q_j\right) \frac{e^{s_i}}{1 + \sum_{k' \in \mathcal{C}_x} e^{s_{k'}} + e^{s_j}/q_j} \right].
\end{aligned}$$

Similarly, for $j \in \mathcal{N}_{\boldsymbol{x}}$, we have

$$\begin{aligned}
\nabla_{s_j} R &= \mathbb{E}_{\boldsymbol{x}}\, q_j \left[-\left(p_K + \sum_{k \in \mathcal{C}_{\boldsymbol{x}}} p_k\right) \frac{e^{s_j}/q_j}{1 + \sum_{k' \in \mathcal{C}_x} e^{s_{k'}} + e^{s_j}/q_j} + p_j/q_j \left(1 - \frac{e^{s_j}/q_j}{1 + \sum_{k' \in \mathcal{C}_x} e^{s_{k'}} + e^{s_j}/q_j}\right) \right] \\
&= \mathbb{E}_{\boldsymbol{x}}\, p_j - \left(p_K + \sum_{k \in \mathcal{C}_{\boldsymbol{x}}} p_k + p_j/q_j\right) \frac{e^{s_j}}{1 + \sum_{k' \in \mathcal{C}_x} e^{s_{k'}} + e^{s_j}/q_j}.
\end{aligned}$$

By measuring $s_k = \log \frac{p_k}{p_K}$, we see that $\nabla_{s_k} R = 0$ for $k = 1, \cdots, K-1$. Therefore, $s_k = \log \frac{p_k}{p_K}$ is an extrema of R. Now, for $i, i' \in C_{\boldsymbol{x}}$ and $j, j' \in \mathcal{N}_{\boldsymbol{x}}$, we have

$$\begin{aligned}
\mathbb{H}_{ii} &= \nabla^2_{s_i s_i} R = -\mathbb{E}_{\boldsymbol{x}} \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_j D_j \frac{e^{s_i}(E_j - e^{s_i})}{E_j^2}, \\
\mathbb{H}_{ii'} &= \nabla^2_{s_i s_{i'}} R = \mathbb{E}_{\boldsymbol{x}} \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_j D_j \frac{e^{s_i} e^{s_{i'}}}{E_j^2}, \\
\mathbb{H}_{ij} &= \mathbb{H}_{ji} = \nabla^2_{s_i s_j} R = \nabla^2_{s_j s_i} R = \mathbb{E}_{\boldsymbol{x}} \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} D_j \frac{e^{s_i} e^{s_j}}{E_j^2}, \\
\mathbb{H}_{jj} &= \nabla^2_{s_j s_j} R = -\mathbb{E}_{\boldsymbol{x}}\, D_j \frac{e^{s_j}(E_j - e^{s_j}/q_j)}{E_j^2}, \\
\mathbb{H}_{jj'} &= \nabla^2_{s_j s_{j'}} R = 0,
\end{aligned}$$

where

$$D_j = p_K + \sum_{k' \in \mathcal{C}_x} p_{k'} + p_j/q_j.$$

Now, we can write

$$\begin{aligned}
\nabla^2_s R &= \left[\begin{array}{ccc|cccccc}
\mathbb{H}_{i_1 i_1} & \cdots & \mathbb{H}_{i_1 i_{|\mathcal{C}_{\boldsymbol{x}}|}} & 0 & \cdots & \mathbb{H}_{i_1 j} & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\mathbb{H}_{i_{|\mathcal{C}_{\boldsymbol{x}}|} i_1} & \cdots & \mathbb{H}_{i_{|\mathcal{C}_{\boldsymbol{x}}|} i_{|\mathcal{C}_{\boldsymbol{x}}|}} & 0 & \cdots & \mathbb{H}_{i_{|\mathcal{C}_{\boldsymbol{x}}|} j} & \cdots & 0 \\
\hline
0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\mathbb{H}_{j i_1} & \cdots & \mathbb{H}_{j i_{|\mathcal{C}_{\boldsymbol{x}}|}} & 0 & \cdots & \mathbb{H}_{jj} & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0
\end{array}\right] \\
&= -\mathbb{E}_{\boldsymbol{x}} \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_j \frac{D_j}{E_j} \left[diag(\boldsymbol{v}_j) - \frac{1}{E_j} \boldsymbol{v}_j \boldsymbol{v}_j^\top \right].
\end{aligned}$$

where $\boldsymbol{v}_j = (e^{s_{i_1}}, \cdots, e^{s_{i_{|\mathcal{C}_{\boldsymbol{x}}|}}}, 0, \cdots, e^{s_j}/q_j, \cdots, 0)^\top$. Let

$$\boldsymbol{A}_j = diag(\boldsymbol{v}_j) - \frac{1}{E_j} \boldsymbol{v}_j \boldsymbol{v}_j^\top.$$

For any non-zero vector $\boldsymbol{\varphi} = (\varphi_1, \cdots, \varphi_{K-1})^\top \in \mathbb{R}^{K-1}$, we have

$$\boldsymbol{\varphi}^\top \boldsymbol{A}_j \boldsymbol{\varphi} = \sum_{i\in\mathcal{C}_{\boldsymbol{x}}} e^{s_i}\varphi_i^2 + \frac{e^{s_j}}{q_j}\varphi_j^2 - \frac{1}{E_j}\left(\sum_{i\in\mathcal{C}_{\boldsymbol{x}}} e^{s_i}\varphi_i + \frac{e^{s_j}}{q_j}\varphi_j\right)^2 \geq \frac{\left(\sum_{i\in\mathcal{C}_{\boldsymbol{x}}} e^{s_i}\varphi_i + \frac{e^{s_j}}{q_j}\varphi_j\right)^2}{\sum_{i\in\mathcal{C}_{\boldsymbol{x}}} e^{s_i} + \frac{e^{s_j}}{q_j}} - \frac{1}{E_j}\left(\sum_{i\in\mathcal{C}_{\boldsymbol{x}}} e^{s_i}\varphi_i + \frac{e^{s_j}}{q_j}\varphi_j\right)^2 > 0,$$

for every $j \in \mathcal{N}_{\boldsymbol{x}}$, where the first inequality is by the Cauchy-Schwarz inequality and the second inequality is because $0 < \sum_{i\in\mathcal{C}_{\boldsymbol{x}}} e^{s_i} + \frac{e^{s_j}}{q_j} < E_j$. Therefore, $-\nabla_s^2 R = \mathbb{E}_{\boldsymbol{x}} \sum_{j\in\mathcal{N}_{\boldsymbol{x}}} q_j \frac{D_j}{E_j}\boldsymbol{A}_j$ is positive-definite and R is strongly concave with respect to s. Hence, $s_k = \log\frac{p_k}{p_K}$ for $k = 1, \cdots, K-1$ is the only maxima of R. □

A.3. Proof of Theorem 2

Proof. R can be re-written as

$$R(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}} \sum_{k\in\mathcal{C}_{\boldsymbol{x}}} p_k \sum_{j\in\mathcal{N}_{\boldsymbol{x}}} q_j \log\frac{e^{s_k}}{E_j} + \sum_{k\in\mathcal{N}_{\boldsymbol{x}}} p_k \log\frac{e^{s_k}}{E_k} + p_K \sum_{j\in\mathcal{N}_{\boldsymbol{x}}} q_j \log\frac{1}{E_j}.$$

Note that E_j for any j can be viewed as a function of $\boldsymbol{s} = (s_1, \cdots, s_{K-1})^\top$. Define the following function

$$G(\boldsymbol{s}) = \sum_{k\in\mathcal{C}_{\boldsymbol{x}}} p_k \sum_{j\in\mathcal{N}_{\boldsymbol{x}}} q_j \log E_j + \sum_{k\in\mathcal{N}_{\boldsymbol{x}}} p_k \log E_k + p_K \sum_{j\in\mathcal{N}_{\boldsymbol{x}}} q_j \log E_j,$$

then for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$,

$$\begin{aligned}
R(\boldsymbol{\theta}^*) - R(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{x}} \sum_{k\in\mathcal{C}_{\boldsymbol{x}}} p_k \sum_{j\in\mathcal{N}_{\boldsymbol{x}}} q_j \left(\log\frac{E_j}{E_j^*} + s_k^* - s_k\right) + \sum_{k\in\mathcal{N}_{\boldsymbol{x}}} p_k \left(\log\frac{E_k}{E_k^*} + s_k^* - s_k\right) + p_K \sum_{j\in\mathcal{N}_{\boldsymbol{x}}} q_j \log\frac{E_j}{E_j^*} \\
&= \mathbb{E}_{\boldsymbol{x}} \sum_{k\in\mathcal{C}_{\boldsymbol{x}}} p_k \sum_{j\in\mathcal{N}_{\boldsymbol{x}}} q_j \log\frac{E_j}{E_j^*} + \sum_{k\in\mathcal{N}_{\boldsymbol{x}}} p_k \log\frac{E_k}{E_k^*} + p_K \sum_{j\in\mathcal{N}_{\boldsymbol{x}}} q_j \log\frac{E_j}{E_j^*} + \sum_{k=1}^{K-1} p_k(s_k^* - s_k) \\
&= G(\boldsymbol{s}) - G(\boldsymbol{s}^*) - \nabla G(\boldsymbol{s}^*)^\top(\boldsymbol{s} - \boldsymbol{s}^*) = \Delta(\boldsymbol{s}, \boldsymbol{s}^*),
\end{aligned}$$

where $\Delta(\boldsymbol{s}, \boldsymbol{s}^*)$ is the Bregman divergence of the convex function $G(\boldsymbol{s})$. Since $G(\cdot)$ is convex, we have $\Delta(s, s^*) \geq 0$ and $\Delta(\boldsymbol{s}, \boldsymbol{s}^*) = 0$ only when $\boldsymbol{s} = \boldsymbol{s}^*$. Under the assumption that the parameter space is compact and $\forall\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$ we have $\mathbb{P}_{\mathcal{X}}\left(s_k(\boldsymbol{x}, \boldsymbol{\theta}) \neq s_k(\boldsymbol{x}, \boldsymbol{\theta}^*)\right) > 0$ for $k \neq K$, we know that $R(\boldsymbol{\theta}) < R(\boldsymbol{\theta}^*)$ for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$.

Given any $\varepsilon' > 0$, there exists $\varepsilon > 0$ that $R(\boldsymbol{\theta}^*) - R(\boldsymbol{\theta}) < \varepsilon$ implies $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\| < \varepsilon'$. Now according to Lemma 1, there exists a $\delta > 0$, when $n \to \infty$, we have

$$\begin{aligned}
R(\boldsymbol{\theta}^*) - R(\hat{\boldsymbol{\theta}}) = R(\boldsymbol{\theta}^*) - \hat{R}_n(\boldsymbol{\theta}^*) + \hat{R}_n(\boldsymbol{\theta}^*) - R(\hat{\boldsymbol{\theta}}) &\leq R(\boldsymbol{\theta}^*) - \hat{R}_n(\boldsymbol{\theta}^*) + \hat{R}_n(\hat{\boldsymbol{\theta}}) - R(\hat{\boldsymbol{\theta}}) \\
&\leq |R(\boldsymbol{\theta}^*) - \hat{R}_n(\boldsymbol{\theta}^*)| + |\hat{R}_n(\hat{\boldsymbol{\theta}}) - R(\hat{\boldsymbol{\theta}})| < 2\delta.
\end{aligned}$$

This implies that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| < \delta'$ for any $\delta' > 0$. □

A.4. Proof of Theorem 3

Proof. By the Mean Value Theorem, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = -\nabla^2 \hat{R}_n(\bar{\boldsymbol{\theta}})^{-1}\sqrt{n}\nabla\hat{R}_n(\boldsymbol{\theta}^*), \tag{12}$$

where $\bar{\boldsymbol{\theta}} = t\boldsymbol{\theta}^* + (1-t)\hat{\boldsymbol{\theta}}$ for some $t \in [0, 1]$. Note that Lemma 1 implies that $\nabla^2\hat{R}_n(\bar{\boldsymbol{\theta}})^{-1}$ converges to $\nabla^2 R(\bar{\boldsymbol{\theta}})^{-1}$ in probability; moreover, $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$ in probability and hence $\bar{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$ in probability. By the Slutsky's Theorem, the limit distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is given by

$$-\nabla^2 R(\boldsymbol{\theta}^*)^{-1}\sqrt{n}\nabla\hat{R}_n(\boldsymbol{\theta}^*).$$

Observe that $\sqrt{n}\nabla\hat{R}_n(\boldsymbol{\theta}^*)$ is the sum of n i.i.d. random vectors with mean $\mathbb{E}\sqrt{n}\nabla\hat{R}_n(\boldsymbol{\theta}^*) = \sqrt{n}\mathbb{E}\nabla R(\boldsymbol{\theta}^*) = 0$, and the variance of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is

$$Var\left(\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right) = \nabla^2 R(\boldsymbol{\theta}^*)^{-1} Var\left(\sqrt{n}\nabla\hat{R}_n(\boldsymbol{\theta}^*)\right)\nabla^2 R(\boldsymbol{\theta}^*)^{-1}.$$

From the proof of Theorem 1, we have

$$\nabla^2 R(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{x}} \boldsymbol{\nabla} \left[\sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_j \frac{D_j}{E_j} \boldsymbol{A}_j \right] \boldsymbol{\nabla}^\top, \tag{13}$$

where

$$\boldsymbol{\nabla} = diag\left(\left(\nabla_{i_1}, \cdots, \nabla_{i_{|\mathcal{C}_{\boldsymbol{x}}|}}, \nabla_{j_1}, \cdots, \nabla_{j_{|\mathcal{N}_{\boldsymbol{x}}|}} \right)^\top \right)$$

and $\nabla_k = \nabla_{\boldsymbol{\theta}} s_k$.
Measuring $\nabla^2 R(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$, we have

$$\nabla^2 R(\boldsymbol{\theta}^*) = -\mathbb{E}_{\boldsymbol{x}} \boldsymbol{\nabla} \boldsymbol{M} \boldsymbol{\nabla}^\top \tag{14}$$

where

$$\boldsymbol{M} = \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_j \left[diag(\boldsymbol{u}_j) - \frac{1}{D_j} \boldsymbol{u}_j \boldsymbol{u}_j^\top \right],$$

where $\boldsymbol{u}_j = (p_{i_1}, \cdots, p_{i_{|\mathcal{C}_{\boldsymbol{x}}|}}, 0, \cdots, p_j/q_j, \cdots, 0)^\top$. By following the proof of Theorem 1, it is easy to show that $\boldsymbol{M} \succ 0$ is positive definite.

Next, we derive $Var\left(\sqrt{n}\nabla \hat{R}_n(\boldsymbol{\theta}^*)\right)$. Introduce some Bernoulli variables Q_j for $j \in \mathcal{N}_{\boldsymbol{x}}$ with $p(Q_j = 1|\boldsymbol{x}) = q_j$. Now, for $i, i' \in C_{\boldsymbol{x}}$ and $j, j' \in \mathcal{N}_{\boldsymbol{x}}$, we have

$$\begin{aligned}
\mathbb{V}_{ii} &= Var\left(\nabla_i \hat{R}_n(\boldsymbol{\theta}^*), \nabla_i \hat{R}_n(\boldsymbol{\theta}^*)\right) \\
&= \mathbb{E}_{\boldsymbol{x},Q}\, Q \left[p_i \left(1 - \frac{e^{s_i^*}}{1 + \sum_{k' \in \mathcal{C}_{\boldsymbol{x}}} e^{s_{k'}^*} + e^{s_j^*}/q_j} \right)^2 + (D_j - p_i) \left(\frac{e^{s_i^*}}{1 + \sum_{k' \in \mathcal{C}_{\boldsymbol{x}}} e^{s_{k'}^*} + e^{s_j^*}/q_j} \right)^2 \right] \cdot \nabla_i \nabla_i^\top \\
&= \mathbb{E}_{\boldsymbol{x}} \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_j \frac{p_i (D_j - p_i)}{D_j} \cdot \nabla_i \nabla_i^\top,
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}_{ii'} &= Var\left(\nabla_i \hat{R}_n(\boldsymbol{\theta}^*), \nabla_{i'} \hat{R}_n(\boldsymbol{\theta}^*)\right) = \mathbb{E}_{\boldsymbol{x},Q}\, Q \left[(D_j - p_i - p_{i'}) \frac{p_i p_{i'}}{D_j^2} - p_i (1 - \frac{p_i}{D_j}) \frac{p_{i'}}{D_j} - p_{i'} (1 - \frac{p_{i'}}{D_j}) \frac{p_i}{D_j} \right] \cdot \nabla_i \nabla_{i'}^\top \\
&= -\mathbb{E}_{\boldsymbol{x}} \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} q_j \frac{p_i p_{i'}}{D_j} \cdot \nabla_i \nabla_{i'}^\top.
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}_{jj} &= Var\left(\nabla_j \hat{R}_n(\boldsymbol{\theta}^*), \nabla_j \hat{R}_n(\boldsymbol{\theta}^*)\right) = \mathbb{E}_{\boldsymbol{x},Q}\, Q \left[\frac{p_j}{q_j} \left(1 - \frac{p_j/q_j}{D_j} \right)^2 + (D_j - p_j/q_j) \frac{p_j^2/q_j^2}{D_j^2} \right] \cdot \nabla_j \nabla_j^\top \\
&= \mathbb{E}_{\boldsymbol{x}} \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} \frac{p_j \left(D_j - p_j/q_j\right)}{D_j} \cdot \nabla_j \nabla_j^\top.
\end{aligned}$$

$$\mathbb{V}_{jj'} = \mathbf{0}.$$

$$\begin{aligned}
\mathbb{V}_{ij} = \mathbb{V}_{ji} &= Var\left(\nabla_i \hat{R}_n(\boldsymbol{\Theta}^*), \nabla_j \hat{R}_n(\boldsymbol{\Theta}^*)\right) \\
&= \mathbb{E}_{\boldsymbol{x},\boldsymbol{Q}}\, Q \left[(D_j - p_i - p_j/q_j) \frac{p_i p_j/q_j}{D_j^2} - p_i \left(1 - \frac{p_i}{D_j} \right) \frac{p_j/q_j}{D_j} - p_j/q_j \left(1 - \frac{p_j/q_j}{D_j} \right) \frac{p_i}{D_j} \right] \cdot \nabla_i \nabla_{i'}^\top \\
&= -\mathbb{E}_{\boldsymbol{x}} \sum_{j \in \mathcal{N}_{\boldsymbol{x}}} \frac{p_i p_j}{D_j} \cdot \nabla_i \nabla_{i'}^\top.
\end{aligned}$$

Now, the variance can be written as

$$V(\boldsymbol{\theta}^*) = Var\left(\sqrt{n}\nabla \hat{R}_n(\boldsymbol{\theta}^*)\right)$$

$$= \left[\begin{array}{ccc|ccccc} \mathbb{V}_{i_1 i_1} & \cdots & \mathbb{V}_{i_1 i_{|\mathcal{C}_{\boldsymbol{x}}|}} & 0 & \cdots & \mathbb{V}_{i_1 j} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbb{V}_{i_{|\mathcal{C}_{\boldsymbol{x}}|} i_1} & \cdots & \mathbb{V}_{i_{|\mathcal{C}_{\boldsymbol{x}}|} i_{|\mathcal{C}_{\boldsymbol{x}}|}} & 0 & \cdots & \mathbb{V}_{i_{|\mathcal{C}_{\boldsymbol{x}}|} j} & \cdots & 0 \\ \hline 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbb{V}_{j i_1} & \cdots & \mathbb{V}_{j i_{|\mathcal{C}_{\boldsymbol{x}}|}} & 0 & \cdots & \mathbb{V}_{jj} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 \end{array}\right].$$

By comparing $\nabla^2 R(\boldsymbol{\theta}^*)$ and $V(\boldsymbol{\theta}^*)$, we immediately have $-\nabla^2 R(\boldsymbol{\theta}^*) = V(\boldsymbol{\theta}^*)$ and hence

$$Var\left(\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right) = \left[\mathbb{E}_{\boldsymbol{x}} \boldsymbol{\nabla} \boldsymbol{M} \boldsymbol{\nabla}^\top\right]^{-1}.$$

□

A.5. Proof of Corollary 1

Proof. By following the proof of Theorem 3, it is easy to show that the statistical variance of the softmax logistic regression in Eq. (1) is $[\mathbb{E}_{\boldsymbol{x}} \boldsymbol{\nabla} \boldsymbol{M}^{mle} \boldsymbol{\nabla}^\top]^{-1}$ (with $s_K = 0$ fixed), where

$$\boldsymbol{M}^{mle} = diag\left(\begin{bmatrix} p_1 \\ \vdots \\ p_{K-1} \end{bmatrix}\right) - \begin{bmatrix} p_1 \\ \vdots \\ p_{K-1} \end{bmatrix} \begin{bmatrix} p_1 \\ \vdots \\ p_{K-1} \end{bmatrix}^\top.$$

When $\sum_{k \in \mathcal{C}_{\boldsymbol{x}} \cup \{K\}} p(k, \boldsymbol{x}) \to 1$, we have $\sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'} \to 0$ and $D_j \to 1$. Then,

$$\boldsymbol{M} = diag\left(\begin{bmatrix} p_{i_1} \\ \vdots \\ p_{i_{|\mathcal{C}_{\boldsymbol{x}}|}} \\ p_{j_1} \\ \vdots \\ p_{j_{|\mathcal{N}_{\boldsymbol{x}}|}} \end{bmatrix}\right) - \left[\begin{array}{ccc|ccc} p_{i_1} p_{i_1} & \cdots & p_{i_1} p_{i_{|\mathcal{C}_{\boldsymbol{x}}|}} & p_{i_1} \sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'} & \cdots & p_{i_1} \sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{i_{|\mathcal{C}_{\boldsymbol{x}}|}} p_{i_1} & \cdots & p_{i_{|\mathcal{C}_{\boldsymbol{x}}|}} p_{i_{|\mathcal{C}_{\boldsymbol{x}}|}} & p_{i_{|\mathcal{C}_{\boldsymbol{x}}|}} \sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'} & \cdots & p_{i_{|\mathcal{C}_{\boldsymbol{x}}|}} \sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'} \\ \hline p_{i_1} \sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'} & \cdots & p_{i_{|\mathcal{C}_{\boldsymbol{x}}|}} \sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'} & p_{j_1}^2 / q_{j_1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{i_1} \sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'} & \cdots & p_{i_{|\mathcal{C}_{\boldsymbol{x}}|}} \sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'} & 0 & \cdots & p_{j_{|\mathcal{N}_{\boldsymbol{x}}|}}^2 / q_{j_{|\mathcal{N}_{\boldsymbol{x}}|}} \end{array}\right].$$

If we arrange the index order in $\boldsymbol{M}^{mle}$ according to the index order in $\boldsymbol{M}$ and denote $\boldsymbol{\Delta} = \boldsymbol{M} - \boldsymbol{M}^{mle}$, we have

$$\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\Delta}_1 & \boldsymbol{\Delta}_2 \\ \boldsymbol{\Delta}_2^\top & \boldsymbol{\Delta}_3 \end{bmatrix} \to \mathbf{0},$$

because

$$\boldsymbol{\Delta}_1 = \mathbf{0},$$

$$\boldsymbol{\Delta}_2 = \begin{bmatrix} p_{i_1}(p_{j_1} - \sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'}) & \cdots & p_{i_1}(p_{j_{|\mathcal{N}_{\boldsymbol{x}}|}} - \sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'}) \\ \cdots & \cdots & \cdots \\ p_{i_{|\mathcal{C}_{\boldsymbol{x}}|}}(p_{j_1} - \sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'}) & \cdots & p_{i_{|\mathcal{C}_{\boldsymbol{x}}|}}(p_{j_{|\mathcal{N}_{\boldsymbol{x}}|}} - \sum_{j' \in \mathcal{N}_{\boldsymbol{x}}} p_{j'}) \end{bmatrix} \to \mathbf{0},$$

$$\boldsymbol{\Delta}_3 = \begin{bmatrix} p_{j_1}^2 (1 - 1/q_{j_1}) & \cdots & p_{j_1} p_{j_{|\mathcal{N}_{\boldsymbol{x}}|}} \\ \cdots & \cdots & \cdots \\ p_{j_{|\mathcal{N}_{\boldsymbol{x}}|}} p_{j_1} & \cdots & p_{j_{|\mathcal{N}_{\boldsymbol{x}}|}}^2 (1 - 1/q_{j_{|\mathcal{N}_{\boldsymbol{x}}|}}) \end{bmatrix} \to \mathbf{0}.$$

This completes the proof. □

B. The Beam Search Algorithm

The beam search algorithm used in both training and testing is depicted in Algorithm 3.

Algorithm 3 The Beam Search Algorithm.

Input: The root of the tree, input data point $\boldsymbol{x}$ and Beam width J.
Output: The J candidate classes.

Initialize stack $\mathcal{S} \leftarrow root$ and stack $\mathcal{S}' \leftarrow \emptyset$;
Initialize the candidate class set $\mathcal{E} \leftarrow \emptyset$;
while true **do**
    if $\mathcal{S}$ is empty **then**
        Break;
    end if
    for $i = 1$ to $\mathcal{S}.size()$ **do**
        if $\mathcal{S}_i$ is a leaf **then**
            $\mathcal{E}.pushback(\mathcal{S}_i)$;
        else
            for $c = 1$ to $\mathcal{S}_i.Child.size()$ **do**
                Accumulate the score to $\mathcal{S}_i.Child(c)$;
                $\mathcal{S}'.pushback(\mathcal{S}_i.Child(c))$;
            end for
        end if
    end for
    $\mathcal{S}.clear()$;
    if $\mathcal{S}'.size() > J$ **then**
        // Using the max heap.
        Find the top-J nodes with the highest accumulated scores in $\mathcal{S}'$ and push them into $\mathcal{S}$;
    else
        $\mathcal{S} \leftarrow \mathcal{S}'$;
    end if
    $\mathcal{S}'.clear()$;
end while
// Using the max heap.
Return the top-J classes with the highest scores in $\mathcal{E}$;

C. A Hierarchical Clustering Method for Generating the Tree Structure

Given the data points of a dataset, we can obtain the center, i.e., the average data point, of each class by scanning the data once and get $\bar{\boldsymbol{X}} \in \mathbb{R}^{K \times d}$, where K is the number of classes and d is the feature dimension. Then, a hierarchical clustering algorithm in Algorithm 4 is performed by viewing each row of $\bar{\boldsymbol{X}}$ as a separate data point. In Algorithm 4, the function 'Split(root)' in step 16 has already constructed a b-nary tree, which can be used by the Beam Tree Algorithm. However, the clustering algorithm, e.g., the k-means algorithm, may generate imbalanced clusters in step 9, and the resulting b-nary tree in step 16 may be imbalanced and affect the efficiency of Beam Tree. A simple way to fix this problem is to fetch the labels (leaves) in the tree in step 16 from left to right, where the obtained label order maintains a rough similarity relationship among the classes. We then assign the ordered labels to the leaves of a new balanced b-nary tree from left to right.

D. Experimental Details

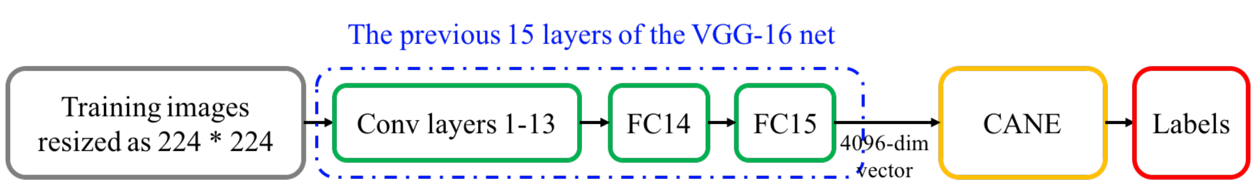

Figure 4. The neural network structure used for the ImageNet datasets. 'FC' indicates fully-connected layer.

Algorithm 4 A Hierarchical Clustering Algorithm for Generating the Tree over Class Labels.

Input: K, b and $\bar{\boldsymbol{X}}$.
Output: a b-nary tree.

Function Split(node o)
while true **do**
  if o is assigned with only one label **then**
    $o.isleaf = true$;
    Return;
  end if
  Perform any clustering algorithm, e.g., k-means, on the labels associated with the node o and obtain b clusters $\{\mathcal{L}_1, \cdots, \mathcal{L}_b\}$;
  Split o into b children $\{o_1, \cdots, o_b\}$ and assign the label clusters $\{\mathcal{L}_1, \cdots, \mathcal{L}_b\}$ to them respectively;
  for $i = 1$ to b **do**
    Split(o_i);
  end for
end while

Assign root with all labels $\{1, 2, \cdots, K\}$;
Split(root);
Get the label order in the leaves from left to right;
Assign the labels to the leaves of a new balanced b-nary tree from left to right;
Return the balanced b-nary tree;

Hyper-parameter tuning is computationally expensive. In order to efficiently select a good setting of the hyper-parameters, we let each method process half epoch of the training data and use another 10% held-out subset of the training set to tune hyper-parameters. For every classifier, the learning rate η needs to be tuned. For the LOMTree method, by following (Choromanska & Langford, 2015), we choose the number of the internal nodes in its binary tree from a set $\{K-1, 4K-1, 16K-1, 64K-1\}$, and tune the swap resistance from $\{4, 16, 64, 256\}$. The Recall Tree method has a default setting for large class problem in (Daume III et al., 2017), which is also adopted in the experiments.

The VGG-16 network structure used in ImageNet-2010 and ImageNet-10K datasets is provided in Fig. 4. Parameters of Conv layers 1-13, FC14 and FC15 are pre-trained on the ImageNet 2012 dataset.