1 Supplementary Material: Complete Proof of Theorem 4

Define two new functions $\widetilde{L}_\lambda$ and $\widetilde{H}_\alpha$ as

$$\widetilde{L}_\lambda := \frac{1}{\lambda - 1}\left[\sum_{j\in\mathcal{L}} \pi_{\Theta_j}\lambda^{d_j} - 1\right] = \sum_{j\in\mathcal{L}} \pi_{\Theta_j}\left[\sum_{h=0}^{d_j-1}\lambda^h\right]$$
$$\widetilde{H}_\alpha := 1 - \frac{1}{\left(\sum_{k=1}^K \pi_{\Theta^k}^\alpha\right)^{\frac{1}{\alpha}}},$$

where $\widetilde{L}_\lambda$ is related to the cost function $L_\lambda(\Pi)$ as

$$\lambda^{L_\lambda(\Pi)} = (\lambda - 1)\widetilde{L}_\lambda + 1, \tag{1}$$

and $\widetilde{H}_\alpha$ is related to the α-Rényi entropy $H_\alpha(\Pi_{\mathbf{y}})$ as

$$H_\alpha(\Pi_{\mathbf{y}}) = \frac{1}{1-\alpha}\log_2 \sum_{k=1}^K \pi_{\Theta^k}^\alpha = \frac{1}{\alpha\log_2\lambda}\log_2 \sum_{k=1}^K \pi_{\Theta^k}^\alpha = \log_\lambda\left(\sum_{k=1}^K \pi_{\Theta^k}^\alpha\right)^{\frac{1}{\alpha}} \tag{2a}$$

$$\implies \lambda^{H_\alpha(\Pi_{\mathbf{y}})} = \left(\sum_{k=1}^K \pi_{\Theta^k}^\alpha\right)^{\frac{1}{\alpha}} = \left(\sum_{k=1}^K \pi_{\Theta^k}^\alpha\right)^{\frac{1}{\alpha}} \widetilde{H}_\alpha + 1 \tag{2b}$$

where we use the definition of α, i.e., $\alpha = \frac{1}{1+\log_2\lambda}$ in (2a).

Now, we note from Lemma 1 that $\widetilde{L}_\lambda$ can be decomposed as

$$\widetilde{L}_\lambda = \sum_{a\in\mathcal{I}} \lambda^{d_a}\pi_{\Theta_a}$$
$$\implies \lambda^{L_\lambda(\Pi)} = 1 + \sum_{a\in\mathcal{I}}(\lambda-1)\lambda^{d_a}\pi_{\Theta_a} \tag{3}$$

where d_a denotes the depth of internal node 'a' in the tree T. Similarly, note from Lemma 2 that $\widetilde{H}_\alpha$ can be decomposed as

$$\widetilde{H}_\alpha = \frac{1}{\left(\sum_{k=1}^K \pi_{\Theta^k}^\alpha\right)^{\frac{1}{\alpha}}}\sum_{a\in\mathcal{I}}\left[\pi_{\Theta_a}\mathcal{D}_\alpha(\Theta_a) - \pi_{\Theta_{l(a)}}\mathcal{D}_\alpha(\Theta_{l(a)}) - \pi_{\Theta_{r(a)}}\mathcal{D}_\alpha(\Theta_{r(a)})\right]$$
$$\implies \lambda^{H_\alpha(\Pi_{\mathbf{y}})} = 1 + \sum_{a\in\mathcal{I}}\left[\pi_{\Theta_a}\mathcal{D}_\alpha(\Theta_a) - \pi_{\Theta_{l(a)}}\mathcal{D}_\alpha(\Theta_{l(a)}) - \pi_{\Theta_{r(a)}}\mathcal{D}_\alpha(\Theta_{r(a)})\right]. \tag{4}$$

Finally, the result follows from (3) and (4) above.

Lemma 1. *The function $\widetilde{L}_\lambda$ can be decomposed over the internal nodes in a tree T, as*

$$\widetilde{L}_\lambda = \sum_{a\in\mathcal{I}} \lambda^{d_a}\pi_{\Theta_a}$$

where d_a denotes the depth of internal node $a \in \mathcal{I}$ and π_{Θ_a} is the probability mass of the objects at that node.

Proof. Let T_a denote a subtree from any internal node 'a' in the tree T and let $\mathcal{I}_a, \mathcal{L}_a$ denote the set of internal nodes and leaf nodes in the subtree T_a, respectively. Then, define $\widetilde{L}_\lambda^a$ in the subtree T_a to be

$$\widetilde{L}_\lambda^a = \sum_{j\in\mathcal{L}_a}\frac{\pi_{\Theta_j}}{\pi_{\Theta_a}}\left[\sum_{h=0}^{d_j^a-1}\lambda^h\right]$$

where d_j^a denotes the depth of leaf node $j \in \mathcal{L}_a$ in the subtree T_a.

Now, we show using induction that for any subtree T_a in the tree T, the following relation holds

$$\pi_{\Theta_a}\widetilde{L}_\lambda^a = \sum_{s\in\mathcal{I}_a} \lambda^{d_s^a}\pi_{\Theta_s} \tag{5}$$

where d_s^a denotes the depth of internal node $s \in \mathcal{I}_a$ in the subtree T_a.

The relation holds trivially for any subtree T_a rooted at an internal node $a \in \mathcal{I}$ whose both child nodes terminate as leaf nodes, with both the left hand side and the right hand side of the expression equal to π_{Θ_a}. Now, consider a subtree T_a rooted at an internal node $a \in \mathcal{I}$ whose left child (or right child) alone terminates as a leaf node. Assume that the above relation holds true for the subtree rooted at the right child of node 'a'. Then,

$$\begin{aligned}
\pi_{\Theta_a}\widetilde{L}_\lambda^a &= \sum_{j\in\mathcal{L}_a} \pi_{\Theta_j}\left[\sum_{h=0}^{d_j^a-1}\lambda^h\right] \\
&= \sum_{\{j\in\mathcal{L}_a:d_j^a=1\}} \pi_{\Theta_j} + \sum_{\{j\in\mathcal{L}_a:d_j^a>1\}} \pi_{\Theta_j}\left[\sum_{h=0}^{d_j^a-1}\lambda^h\right] \\
&= \pi_{\Theta_{l(a)}} + \sum_{\{j\in\mathcal{L}_a:d_j^a>1\}} \pi_{\Theta_j}\left[1+\lambda\sum_{h=0}^{d_j^a-2}\lambda^h\right] \\
&= \pi_{\Theta_a} + \lambda\sum_{j\in\mathcal{L}_{r(a)}} \pi_{\Theta_j}\left[\sum_{h=0}^{d_j^{r(a)}-1}\lambda^h\right] \\
&= \pi_{\Theta_a} + \lambda\sum_{s\in\mathcal{I}_{r(a)}} \lambda^{d_s^{r(a)}}\pi_{\Theta_s}
\end{aligned}$$

where the last step follows from the induction hypothesis. Finally, consider a subtree T_a rooted at an internal node $a \in \mathcal{I}$ whose neither child node terminates as a leaf node. Assume that the relation in (5) holds true for the subtrees rooted at its left and right child nodes. Then,

$$\begin{aligned}
\pi_{\Theta_a}\widetilde{L}_\lambda^a &= \sum_{j\in\mathcal{L}_a} \pi_{\Theta_j}\left[\sum_{h=0}^{d_j^a-1}\lambda^h\right] \\
&= \sum_{j\in\mathcal{L}_{l(a)}} \pi_{\Theta_j}\left[1+\lambda\sum_{h=0}^{d_j^a-2}\lambda^h\right] + \sum_{j\in\mathcal{L}_{r(a)}} \pi_{\Theta_j}\left[1+\lambda\sum_{h=0}^{d_j^a-2}\lambda^h\right] \\
&= \pi_{\Theta_a} + \lambda\sum_{j\in\mathcal{L}_{l(a)}} \pi_{\Theta_j}\left[\sum_{h=0}^{d_j^{l(a)}-1}\lambda^h\right] + \lambda\sum_{j\in\mathcal{L}_{r(a)}} \pi_{\Theta_j}\left[\sum_{h=0}^{d_j^{r(a)}-1}\lambda^h\right] \\
&= \pi_{\Theta_a} + \lambda\left[\sum_{s\in\mathcal{I}_{l(a)}} \lambda^{d_s^{l(a)}}\pi_{\Theta_s} + \sum_{s\in\mathcal{I}_{r(a)}} \lambda^{d_s^{r(a)}}\pi_{\Theta_s}\right] = \sum_{s\in\mathcal{I}_a}\lambda^{d_s^a}\pi_{\Theta_s}
\end{aligned}$$

thereby completing the induction. Finally, the result follows by applying the relation in (5) to the tree T whose probability mass at the root node, $\pi_{\Theta_a} = 1$. $\square$

Lemma 2. *The function $\widetilde{H}_\alpha$ can be decomposed over the internal nodes in a tree T, as*

$$\widetilde{H}_\alpha = \frac{1}{\left(\sum_{k=1}^K \pi_{\Theta^k}^\alpha\right)^{\frac{1}{\alpha}}} \sum_{a\in\mathcal{I}} \left[\pi_{\Theta_a}\mathcal{D}_\alpha(\Theta_a) - \pi_{\Theta_{l(a)}}\mathcal{D}_\alpha(\Theta_{l(a)}) - \pi_{\Theta_{r(a)}}\mathcal{D}_\alpha(\Theta_{r(a)})\right]$$

where $\mathcal{D}_\alpha(\Theta_a) := \left[\sum_{k=1}^K \left(\frac{\pi_{\Theta_a^k}}{\pi_{\Theta_a}}\right)^\alpha\right]^{\frac{1}{\alpha}}$ and π_{Θ_a} denotes the probability mass of the objects at any internal node $a \in \mathcal{I}$.

Proof. Let T_a denote a subtree from any internal node 'a' in the tree T and let $\mathcal{I}_a$ denote the set of internal nodes in the subtree T_a. Then, define $\widetilde{H}^a_\alpha$ in a subtree T_a to be

$$\widetilde{H}^a_\alpha = 1 - \frac{\pi_{\Theta_a}}{\left[\sum_{k=1}^K \pi^\alpha_{\Theta^i_a}\right]^{\frac{1}{\alpha}}}$$

Now, we show using induction that for any subtree T_a in the tree T, the following relation holds

$$\left[\sum_{k=1}^K \pi^\alpha_{\Theta^k_a}\right]^{\frac{1}{\alpha}} \widetilde{H}^a_\alpha = \sum_{s \in \mathcal{I}_a} \left[\pi_{\Theta_s}\mathcal{D}_\alpha(\Theta_s) - \pi_{\Theta_{l(s)}}\mathcal{D}_\alpha(\Theta_{l(s)}) - \pi_{\Theta_{r(s)}}\mathcal{D}_\alpha(\Theta_{r(s)})\right] \tag{6}$$

Note that the relation holds trivially for any subtree T_a rooted at an internal node $a \in \mathcal{I}$ whose both child nodes terminate as leaf nodes. Now, consider a subtree T_a rooted at any other internal node $a \in \mathcal{I}$. Assume the above relation holds true for the subtrees rooted at its left and right child nodes. Then,

$$\begin{aligned}
\left[\sum_{k=1}^K \pi^\alpha_{\Theta^k_a}\right]^{\frac{1}{\alpha}} \widetilde{H}^a_\alpha &= \left[\sum_{k=1}^K \pi^\alpha_{\Theta^k_a}\right]^{\frac{1}{\alpha}} - \pi_{\Theta_a} = \left[\sum_{k=1}^K \pi^\alpha_{\Theta^k_a}\right]^{\frac{1}{\alpha}} - \pi_{\Theta_{l(a)}} - \pi_{\Theta_{r(a)}} \\
&= \left[\sum_{k=1}^K \pi^\alpha_{\Theta^k_a}\right]^{\frac{1}{\alpha}} - \left[\sum_{k=1}^K \pi^\alpha_{\Theta^k_{l(a)}}\right]^{\frac{1}{\alpha}} - \left[\sum_{k=1}^K \pi^\alpha_{\Theta^k_{r(a)}}\right]^{\frac{1}{\alpha}} \\
&\quad + \left(\left[\sum_{k=1}^K \pi^\alpha_{\Theta^k_{l(a)}}\right]^{\frac{1}{\alpha}} - \pi_{\Theta_{l(a)}}\right) + \left(\left[\sum_{k=1}^K \pi^\alpha_{\Theta^k_{r(a)}}\right]^{\frac{1}{\alpha}} - \pi_{\Theta_{r(a)}}\right) \\
&= \left[\pi_{\Theta_a}\mathcal{D}_\alpha(\Theta_a) - \pi_{\Theta_{l(a)}}\mathcal{D}_\alpha(\Theta_{l(a)}) - \pi_{\Theta_{r(a)}}\mathcal{D}_\alpha(\Theta_{r(a)})\right] \\
&\quad + \left[\sum_{k=1}^K \pi^\alpha_{\Theta^k_{l(a)}}\right]^{\frac{1}{\alpha}} \widetilde{H}^{l(a)}_\alpha + \left[\sum_{k=1}^K \pi^\alpha_{\Theta^k_{r(a)}}\right]^{\frac{1}{\alpha}} \widetilde{H}^{r(a)}_\alpha \\
&= \sum_{s \in \mathcal{I}_a} \left[\pi_{\Theta_s}\mathcal{D}_\alpha(\Theta_s) - \pi_{\Theta_{l(s)}}\mathcal{D}_\alpha(\Theta_{l(s)}) - \pi_{\Theta_{r(s)}}\mathcal{D}_\alpha(\Theta_{r(s)})\right]
\end{aligned}$$

where the last step follows from the induction hypothesis. Finally, the result follows by applying the relation in (6) to the tree T. $\square$

2 Proof of Theorem 3

The result in Theorem 3 follows from the above result where each group is of size one, thereby reducing $\mathcal{D}_\alpha(\Theta_a)$ to

$$\mathcal{D}_\alpha(\Theta_a) = \left[\sum_{i=1}^M \left(\frac{\pi_i \mathbb{I}_{\{\theta_i \in\, \Theta_a\}}}{\pi_{\Theta_a}}\right)^\alpha\right]^{\frac{1}{\alpha}} = \left[\sum_{\{i:\theta_i \in\, \Theta_a\}} \left(\frac{\pi_i}{\pi_{\Theta_a}}\right)^\alpha\right]^{\frac{1}{\alpha}},$$

where $\mathbb{I}_{\{\theta_i \in\, \Theta_a\}}$ is the indicator function which takes the value one when $\theta_i \in \Theta_a$, and zero otherwise.

3 Proof of Theorem 2

The result in Theorem 2 is a special case of that in Theorem 4 when $\lambda \to 1$. It follows by taking the logarithm to the base λ on both sides of equation

$$\lambda^{L_\lambda(\Pi)} = \lambda^{H_\alpha(\Pi_{\mathbf{y}})} + \sum_{a \in \mathcal{I}} \pi_{\Theta_a}\left[(\lambda - 1)\lambda^{d_a} - \mathcal{D}_\alpha(\Theta_a) + \frac{\pi_{\Theta_{l(a)}}}{\pi_{\Theta_a}}\mathcal{D}_\alpha(\Theta_{l(a)}) + \frac{\pi_{\Theta_{r(a)}}}{\pi_{\Theta_a}}\mathcal{D}_\alpha(\Theta_{r(a)})\right],$$

and then finding the limit as $\lambda \to 1$.

Using L'Hôpital's rule, the left hand side (LHS) of the equation reduces to

$$\lim_{\lambda \to 1} \log_\lambda(\text{LHS}) = \lim_{\lambda \to 1} L_\lambda(\Pi) = \sum_{j \in \mathcal{L}} \pi_{\Theta_j} d_j,$$

where $L_\lambda(\Pi) = \log_\lambda \left(\sum_{j \in \mathcal{L}} \pi_{\Theta_j} \lambda^{d_j} \right)$. Similarly, the right hand side (RHS) of the equation reduces to

$$\lim_{\lambda \to 1} \log_\lambda(\text{RHS}) = H(\Pi_{\mathbf{y}}) + \sum_{a \in \mathcal{I}} \pi_{\Theta_a} \left[1 - \left(H(\Theta_a) - \frac{\pi_{\Theta_{l(a)}}}{\pi_{\Theta_a}} H(\Theta_{l(a)}) - \frac{\pi_{\Theta_{r(a)}}}{\pi_{\Theta_a}} H(\Theta_{r(a)}) \right) \right],$$

where $H(\Theta_a) = -\sum_{k=1}^{K} \frac{\pi_{\Theta_a^k}}{\pi_{\Theta_a}} \log_2 \left(\frac{\pi_{\Theta_a^k}}{\pi_{\Theta_a}} \right)$.

Finally, the result follows by noticing that

$$\begin{aligned}
& H(\Theta_a) - \frac{\pi_{\Theta_{l(a)}}}{\pi_{\Theta_a}} H(\Theta_{l(a)}) - \frac{\pi_{\Theta_{r(a)}}}{\pi_{\Theta_a}} H(\Theta_{r(a)}) \\
&\quad = \frac{1}{\pi_{\Theta_a}} \left[\sum_{k=1}^{K} \pi_{\Theta_a^k} \log_2 \left(\frac{\pi_{\Theta_a}}{\pi_{\Theta_a^k}} \right) - \pi_{\Theta_{l(a)}^k} \log_2 \left(\frac{\pi_{\Theta_{l(a)}}}{\pi_{\Theta_{l(a)}^k}} \right) - \pi_{\Theta_{r(a)}^k} \log_2 \left(\frac{\pi_{\Theta_{r(a)}}}{\pi_{\Theta_{r(a)}^k}} \right) \right] && \text{(7a)} \\
&\quad = \frac{1}{\pi_{\Theta_a}} \left[\sum_{k=1}^{K} \pi_{\Theta_{l(a)}^k} \log_2 \left(\frac{\pi_{\Theta_a}}{\pi_{\Theta_{l(a)}}} \cdot \frac{\pi_{\Theta_{l(a)}^k}}{\pi_{\Theta_a^k}} \right) + \pi_{\Theta_{r(a)}^k} \log_2 \left(\frac{\pi_{\Theta_a}}{\pi_{\Theta_{r(a)}}} \cdot \frac{\pi_{\Theta_{r(a)}^k}}{\pi_{\Theta_a^k}} \right) \right] && \text{(7b)} \\
&\quad = \frac{1}{\pi_{\Theta_a}} \left[\pi_{\Theta_{l(a)}} \log_2 \left(\frac{\pi_{\Theta_a}}{\pi_{\Theta_{l(a)}}} \right) + \pi_{\Theta_{r(a)}} \log_2 \left(\frac{\pi_{\Theta_a}}{\pi_{\Theta_{r(a)}}} \right) \right. \\
&\qquad\qquad \left. + \sum_{k=1}^{K} \pi_{\Theta_{l(a)}^k} \log_2 \left(\frac{\pi_{\Theta_{l(a)}^k}}{\pi_{\Theta_a^k}} \right) + \pi_{\Theta_{r(a)}^k} \log_2 \left(\frac{\pi_{\Theta_{r(a)}^k}}{\pi_{\Theta_a^k}} \right) \right] && \text{(7c)} \\
&\quad = H(\rho_a) + \sum_{k=1}^{K} \frac{\pi_{\Theta_a^k}}{\pi_{\Theta_a}} H(\rho_a^k), && \text{(7d)}
\end{aligned}$$

where (7b) follows from (7a) by using the relation $\pi_{\Theta_a^k} = \pi_{\Theta_{l(a)}^k} + \pi_{\Theta_{r(a)}^k}$, and (7d) follows from (7c) using the definitions of ρ_a and ρ_a^k.

4 Proof of Theorem 1

The result in Theorem 1 follows from the above result where each group is of size one, thereby having $\rho_a^k = 1 \; \forall k$ at each internal node $a \in \mathcal{I}$. It can also be derived as the limiting case of the relation in Theorem 3 by taking logarithm to the base λ on both sides of the relation and letting $\lambda \to 1$.