E The NP-hardness on optimality of improved coding method

In section 4.2.1, we present that the code length can be further improved by entropy coding. This section will present that if we replace the original code length in problem 1 with the improved code length, the corresponding problem is still NP-hard.

We first introduce the notations that will be used in this section. For a binary tree T, **we denote the set of all internal nodes of T as Int(T).** For $v \in \mathrm{Int}(T)$ and $i \in \{0,1,2\}$, let $f_i(v)$ be the total number of symbol i that produced by v among the dataset $\mathcal{S}$. By the entropy coding, the code length produced by v is

$$L^*(v) = \sum_{i\in\{0,1,2\}} f_i(v)\log_2\left(\frac{\sum_j f_j(v)}{f_i(v)}\right).$$

Thus, the total code length by the improved method for a given tree is

$$L^*(\mathcal{S}) = \sum_{v\in\mathrm{Int}(T)} L^*(v)$$

Problem 3. *Input a dataset $\mathcal{S} = \{S_1, S_2, \ldots, S_N\}$, a binary tree T with K leaf nodes and a number L. Decide if there exists an assignment of K items to the K leaf nodes, such that*

$$\sum_{v\in Int(T)} L^*(v) \le L$$

Theorem 3. *Problem 3 is NP-hard.*

E.1 The NP-hardness of problem 3

Our proof is also based on a reduction from 3-dimensional matching. Let the input of 3-dimensional matching be three disjoint nodes set X, Y, and Z, where $|X| = |Y| = |Z| = k$. The input hyper-edge set is $E \subseteq X \times Y \times Z$. The decision problem of 3-dimensional matching is inputting E and a number s, deciding whether a matching $M \subseteq E$ exists such that $|M| \ge s$.

Without loss of generality, we can assume that k is a power of 2. Since when $k < 2^l < 2k$, we can add $2^l - k$ dummy nodes to each node set X, Y and Z. Then the output of the 3-dimensional matching problem for our new X, Y and Z remains the same as the original one. In the following context, we assume $k = 2^l$.

The goal is to show that we can construct an input of Problem 3 to solve the 3-dimensional matching problem.

First, we build the 4-dimensional matching problem which is equivalent to the original 3-dimensional matching problem. We add an additional node set W of size k. We build a new hyper-edge set $E' \subseteq X \times Y \times Z \times W$ as follows.

$$E' = \left\{(x,y,z,w)\middle|(x,y,z) \in E, w \in W\right\}$$

Then we have the 4-dimensional matching problem that decides whether a matching $M' \subseteq E'$ exists such that $|M'| \ge s$. We can easily see the following result.

Lemma 2. *The corresponding 4-dimensional matching problem has the same output as the original 3-dimensional matching problem.*

Proof. **($\Rightarrow$)** Suppose there exists a matching $M' \subseteq E'$ such that $|M'| \ge s$. Consider the set $M = \{(x,y,z)|(x,y,z,w) \in M'\}$. Since for any $(x,y,z,w) \in M'$, $(x,y,z,w) \in E'$, we have $(x,y,z) \in E$. Therefore, we have $M \subseteq E$.

Since M' is a matching, for any $(x,y,z,w), (x',y',z',w') \in M'$, we have $x \ne x'$, $y \ne y'$, $z \ne z'$ and $w \ne w'$. Thus, for any $(x,y,z), (x',y',z') \in M$, we have $x \ne x'$, $y \ne y'$, and $w \ne w'$. We have that M is a matching.

Thus, there exists a matching $M \subseteq E$ such that $|M| \ge s$.

($\Leftarrow$) On the other hand, assume there exists a matching $M \subseteq E$ such that $|M| \ge s$. Suppose $M = \{(x_1,y_1,z_1), \ldots, (x_m,y_m,z_m)\}$. Since M is a matching, $|M| = m \le k = |W|$. Let $w_1, \ldots, w_m$ be arbitrary m distinct nodes in W. Then $M' = \{(x_1,y_1,z_1,w_1), \ldots, (x_m,y_m,z_m,w_m)\} \subseteq E'$ is a 4-dimensional matching such that $|M'| \ge s$. ∎

Now we construct an input of Problem 3 to solve the 4-dimensional matching problem.

For the item set V, we let $V = X \cup Y \cup Z \cup W$ and we have $|V| = 4k = 2^{l+2}$. For the dataset $\mathcal{S}$, we construct it by the following steps.

(1) $\mathcal{D} = \left\{S = V - \{x,y,z,w\}\middle|(x,y,z,w) \in E'\right\}$.
We add five $S = V$ to $\mathcal{D}$.
(2) $\mathcal{F}$ is initialized with $\mathcal{D}$.
For every two distinct items $a, b \in V$, let

$$\mathrm{MissCount}(a,b;\mathcal{D}) = \sum_{S\in\mathcal{D}} \mathbf{I}[\{a,b\} \subseteq V - S]$$

where $\mathbf{I}[\cdot]$ is the indicator function.
We add $(|\mathcal{D}| - \mathrm{MissCount}(a,b;\mathcal{D}))$ numbers of $S = V - \{a,b\}$ to $\mathcal{F}$.
(3) $\mathcal{S}$ is initialized with $\mathcal{F}$.
For every items $a \in V$, let

$$\mathrm{MissCount}(a;\mathcal{F}) = \sum_{S\in\mathcal{F}} \mathbf{I}[\{a\} \subseteq V - S]$$

We add $(|\mathcal{F}| - \mathrm{MissCount}(a;\mathcal{F}))$ numbers of $S = V - \{a\}$ to $\mathcal{S}$.
We add $|\mathcal{D}|^3$ numbers of $S = V$ to $\mathcal{S}$.

For the construction above, we have the following claims hold:

Claim 5. *For any item $a \in V$,*

$$\mathit{MissCount}(a;\mathcal{S}) = |\mathcal{F}|$$

Proof. If $a \ne b$, then the set $S = V - \{b\}$ will not contribute to the $\mathrm{MissCount}(a;\mathcal{S})$.

Thus,

$$\begin{aligned}&\mathrm{MissCount}(a;\mathcal{S})\\ =&\mathrm{MissCount}(a;\mathcal{F}) + (|\mathcal{F}| - \mathrm{MissCount}(a;\mathcal{F}))\\ =&|\mathcal{F}|\end{aligned}$$

∎

Claim 6. *For any two distinct item $a, b \in V$,*

$$\mathit{MissCount}(a,b;\mathcal{S}) = |\mathcal{D}|$$

Proof. For the sets S whose size is greater than $|V| - 2$, it will not contribute to $\mathrm{MissCount}(a,b;\mathcal{S})$. And if $\{a,b\} \ne \{c,d\}$, the set $S = V - \{c,d\}$ will not contribute to $\mathrm{MissCount}(a,b;\mathcal{S})$.

Thus,

$$\begin{aligned}&\text{MissCount}(a,b;\mathcal{S})\\=&\text{MissCount}(a,b;\mathcal{F})\\=&\text{MissCount}(a,b;\mathcal{D})+(|\mathcal{D}|-\text{MissCount}(a,b;\mathcal{D}))\\=&|\mathcal{D}|\end{aligned}$$

■

Definition 1. *We say $\{a,b,c,d\}$ hits E', if there exist $(x,y,z,w)\in E'$ such that $\{a,b,c,d\}=\{x,y,z,w\}$.*

Claim 7. *For any four distinct items $a,b,c,d\in V$, let*

$$\textit{MissCount}(a,b,c,d;\mathcal{S})=\sum_{S\in\mathcal{S}}\mathbf{I}[\{a,b,c,d\}\subseteq V-S]$$

We have

$$\textit{MissCount}(a,b,c,d;\mathcal{S})=\begin{cases}1, \textit{ if } \{a,b,c,d\} \textit{ hits } E'\\ 0, \textit{ otherwise}\end{cases}$$

Proof. For the sets S of size greater than $|V|-4$ we have $\{a,b,c,d\}\cap S\neq\emptyset$. For those sets $S=V-\{x,y,z,w\}$ that were added in step (1), $\{a,b,c,d\}\cap S\subseteq V-S$ if and only if $\{a,b,c,d\}=\{x,y,z,w\}$. Since the hyper-edges in E are distinct, we have this claim hold. ■

Now, we introduce the tree we input to the problem 3. Let T be a perfect binary tree where each internal node has two children and each leaf node has the same depth. We set the depth of T to be $l+2$ and there are 2^{l+2} leaf nodes in T. Fig. 8 shows an example of the input prefect tree with $2^{l+2}=32$ leaf nodes.

We divide the internal nodes of the perfect tree T into groups.

- We say v is a β-type node if the two children of v are leaf nodes.
- We say v is a α-type node if the two children of v are β-type nodes.
- We say v is a γ-type node if the two children of v are α-type nodes.
- The rest of the internal nodes are called δ-type nodes.

The code length can be decomposed as

$$\begin{aligned}L^*(\mathcal{S})&=\sum_{v\in\text{Int}(T)}L^*(v)\\&=\sum_{\beta\text{-type}v}L^*(v)+\sum_{\alpha\text{-type}v}L^*(v)+\sum_{\gamma\text{-type}v}L^*(v)+\sum_{\delta\text{-type}v}L^*(v)\end{aligned}$$

To understand the above code length, let's examine each term.

Claim 8. *For a β-type node v, we have*

$$L^*(v)=(N+|\mathcal{D}|-2|\mathcal{F}|)\log\frac{N-|\mathcal{D}|}{N+|\mathcal{D}|-2|\mathcal{F}|}+2(|\mathcal{F}|-|\mathcal{D}|)\log\frac{N-|\mathcal{D}|}{|\mathcal{F}|-|\mathcal{D}|}$$

Proof. Suppose a and b are the assigned items of v's left and right child respectively. We have

$$\begin{aligned}f_0(v)&=\sum_{S\in\mathcal{S}}\mathbf{I}[a\in S]\cdot\mathbf{I}[b\notin S]\\f_1(v)&=\sum_{S\in\mathcal{S}}\mathbf{I}[a\notin S]\cdot\mathbf{I}[b\in S]\\f_2(v)&=\sum_{S\in\mathcal{S}}\mathbf{I}[a\in S]\cdot\mathbf{I}[b\in S]\end{aligned}$$

We denote $f_3(v)=\sum_{S\in\mathcal{S}}\mathbf{I}[a\notin S]\cdot\mathbf{I}[b\notin S]$.

Let $|\mathcal{S}|=N$, we have

$$f_0(v)+f_1(v)+f_2(v)+f_3(v)=N$$

By Claim 6, we have

$$f_3(v)=\sum_{S\in\mathcal{S}}\mathbf{I}[\{a,b\}\subseteq V-S]=\text{MissCount}(a,b;\mathcal{S})=|\mathcal{D}|$$

By Claim 5, we have

$$\begin{aligned}f_0(v)+f_3(v)&=\sum_{S\in\mathcal{S}}\mathbf{I}[b\notin S]=\text{MissCount}(b;\mathcal{S})=|\mathcal{F}|\\f_1(v)+f_3(v)&=\sum_{S\in\mathcal{S}}\mathbf{I}[a\notin S]=\text{MissCount}(a;\mathcal{S})=|\mathcal{F}|\end{aligned}$$

By solving those equations, we have

$$\begin{aligned}f_0(v)&=|\mathcal{F}|-|\mathcal{D}|\\f_1(v)&=|\mathcal{F}|-|\mathcal{D}|\\f_2(v)&=N+|\mathcal{D}|-2|\mathcal{F}|\end{aligned}$$

Thus, we have

$$\begin{aligned}&L^*(v)\\=&f_0(v)\log\frac{\sum_{j=0}^2 f_j(v)}{f_0(v)}+f_1(v)\log\frac{\sum_{j=0}^2 f_j(v)}{f_1(v)}+f_2(v)\log\frac{\sum_{j=0}^2 f_j(v)}{f_2(v)}\\=&(N+|\mathcal{D}|-2|\mathcal{F}|)\log\frac{N-|\mathcal{D}|}{N+|\mathcal{D}|-2|\mathcal{F}|}+2(|\mathcal{F}|-|\mathcal{D}|)\log\frac{N-|\mathcal{D}|}{|\mathcal{F}|-|\mathcal{D}|}\end{aligned}$$

■

Claim 9. *For a δ-type node v,*

$$L^*(v)=0$$

Proof. Let u and w be the left and right child of v. Since $|C(u)|=|C(w)|\geq 8$ and $|V-S|\leq 4$ for any $S\in\mathcal{S}$, the v will always output code 2 for every set in $\mathcal{S}$. Thus, $L^*(v)=0$. ■

Claim 10. *Let v be a α-type node. Let u and w be the left and right child of v. Let $C(u)=\{a,b\}$, $C(w)=\{c,d\}$.*

When $\{a,b,c,d\}$ hits E',

$$L^*(v)=(N+1-2|\mathcal{D}|)\log\frac{N-1}{N+1-2|\mathcal{D}|}+2(|\mathcal{D}|-1)\log\frac{N-1}{|\mathcal{D}|-1}.$$

When $\{a,b,c,d\}$ does not hit E',

$$L^*(v)=(N-2|\mathcal{D}|)\log\frac{N}{N-2|\mathcal{D}|}+2|\mathcal{D}|\log\frac{N}{|\mathcal{D}|}$$

Proof. By definition, we have

$$\begin{aligned}f_0(v)&=\sum_{S\in\mathcal{S}}\mathbf{I}[\{a,b\}\nsubseteq V-S]\cdot\mathbf{I}[\{c,d\}\subseteq V-S]\\f_1(v)&=\sum_{S\in\mathcal{S}}\mathbf{I}[\{a,b\}\subseteq V-S]\cdot\mathbf{I}[\{c,d\}\nsubseteq V-S]\\f_2(v)&=\sum_{S\in\mathcal{S}}\mathbf{I}[\{a,b\}\nsubseteq V-S]\cdot\mathbf{I}[\{c,d\}\nsubseteq V-S]\end{aligned}$$

We denote $f_3(v)=\sum_{S\in\mathcal{S}}\mathbf{I}[\{a,b\}\subseteq V-S]\cdot\mathbf{I}[\{c,d\}\subseteq V-S]$.

Let $|\mathcal{S}|=N$, we have

$$f_0(v)+f_1(v)+f_2(v)+f_3(v)=N$$

By Claim 6, we have

$$f_1(v)+f_3(v)=\sum_{S\in\mathcal{S}}\mathbf{I}[\{a,b\}\subseteq V-S]=\text{MissCount}(a,b;\mathcal{S})=|\mathcal{D}|$$

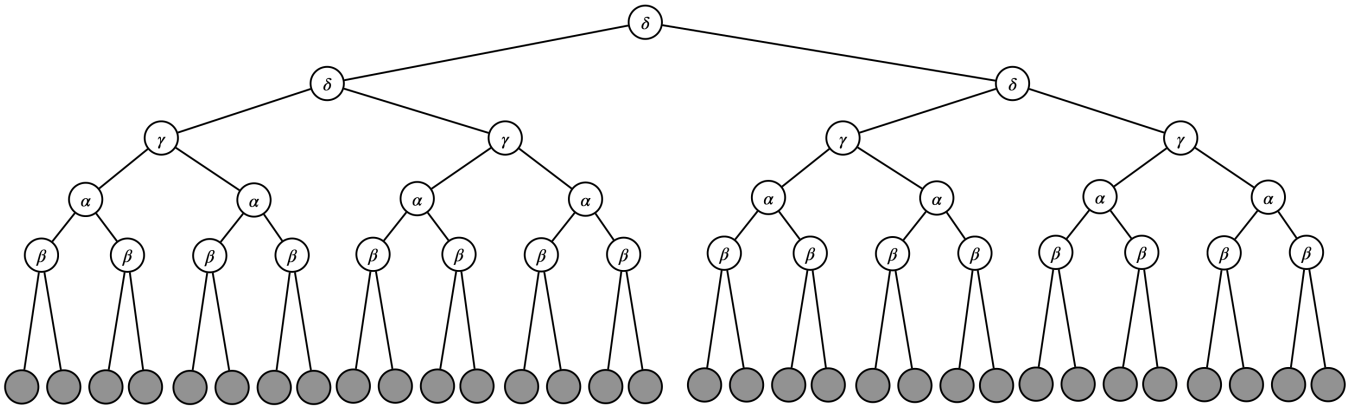


Figure 8: An illustration of the input prefect tree for problem 3.

$$f_0(v) + f_3(v) = \sum_{S\in\mathcal{S}} \mathbf{I}[\{c,d\} \subseteq V - S] = \text{MissCount}(c,d;\mathcal{S}) = |\mathcal{D}|$$

(Case 1) When $\{a,b,c,d\}$ hits E', by Claim 7, we have

$$f_3(v) = \sum_{S\in\mathcal{S}} \mathbf{I}[\{a,b,c,d\} \subseteq V - S] = \text{MissCount}(a,b,c,d;\mathcal{S}) = 1$$

By solving those equations, we have

$$f_0(v) = |\mathcal{D}| - 1$$
$$f_1(v) = |\mathcal{D}| - 1$$
$$f_2(v) = N + 1 - 2|\mathcal{D}|$$

In this case,

$$L^*(v) = (N+1-2|\mathcal{D}|)\log\frac{N-1}{N+1-2|\mathcal{D}|} + 2(|\mathcal{D}|-1)\log\frac{N-1}{|\mathcal{D}|-1}$$

(Case 2) When $\{a,b,c,d\}$ does not hit E', by Claim 7, we have

$$f_3(v) = \sum_{S\in\mathcal{S}} \mathbf{I}[\{a,b,c,d\} \subseteq V - S] = \text{MissCount}(a,b,c,d;\mathcal{S}) = 0$$

By solving those equations, we have

$$f_0(v) = |\mathcal{D}|$$
$$f_1(v) = |\mathcal{D}|$$
$$f_2(v) = N - 2|\mathcal{D}|$$

In this case,

$$L^*(v) = (N-2|\mathcal{D}|)\log\frac{N}{N-2|\mathcal{D}|} + 2|\mathcal{D}|\log\frac{N}{|\mathcal{D}|}$$

■

Claim 11. *Let v be a γ-type node. Let u and w be the left and right child of v.*

When both $C(u)$ and $C(w)$ hits E',

$$L^*(v) = (N-2)\log\frac{N}{N-2} + 2\log N$$

When only one of $C(u)$ and $C(w)$ hits E',

$$L^*(v) = (N-1)\log\frac{N}{N-1} + \log N$$

When none of $C(u)$ and $C(w)$ hits E',

$$L^*(v) = 0$$

Proof. When both $C(u)$ and $C(w)$ hits E', we have

$$f_0(v) = 1, f_1(v) = 1, f_2(v) = N-2$$

Thus, $L^*(v) = (N-2)\log\frac{N}{N-2} + 2\log N$.

When only one of $C(u)$ and $C(w)$ hits E', we have

$$f_0(v) = 0, f_1(v) = 1, f_2(v) = N-1$$

or

$$f_0(v) = 1, f_1(v) = 0, f_2(v) = N-1$$

Thus, $L^*(v) = (N-1)\log\frac{N}{N-1} + \log N$.

When none of $C(u)$ and $C(w)$ hits E', we have

$$f_0(v) = 0, f_1(v) = 0, f_2(v) = N$$

Thus, $L^*(v) = 0$. ■

The following three lemmas state the code length when different hits happen. Fig. 9 gives an overview of the lemmas.

Lemma 3. *Let v and v' be two γ-type nodes. The left and right child of v are u and w. The left and right child of v' are u' and w'. Suppose that both of $C(u)$ and $C(w)$ hits E' and none of $C(u')$ and $C(w')$ hits E'. If we change the items of $C(w)$ with the items of $C(w')$, then $L^*(v) + L^*(v')$ will increase.*

Proof. The original $L^*(v) + L^*(v')$ is

$$(N-2)\log\frac{N}{N-2} + 2\log N + 0.$$

The $L^*(v) + L^*(v')$ after we change the $C(w)$ and $C(w')$ is

$$2 * ((N-1)\log\frac{N}{N-1} + \log N)$$

Take the difference, we have

$$\Delta L = 2(N-1)\log(N-1) - (N-2)\log(N-2) - N\log N$$

By the convexity of $f(x) = x\log x$, we have

$$\frac{1}{2}(f(N) + f(N-2)) > f(\frac{N+(N-2)}{2})$$

Thus, $\Delta L < 0$. ■

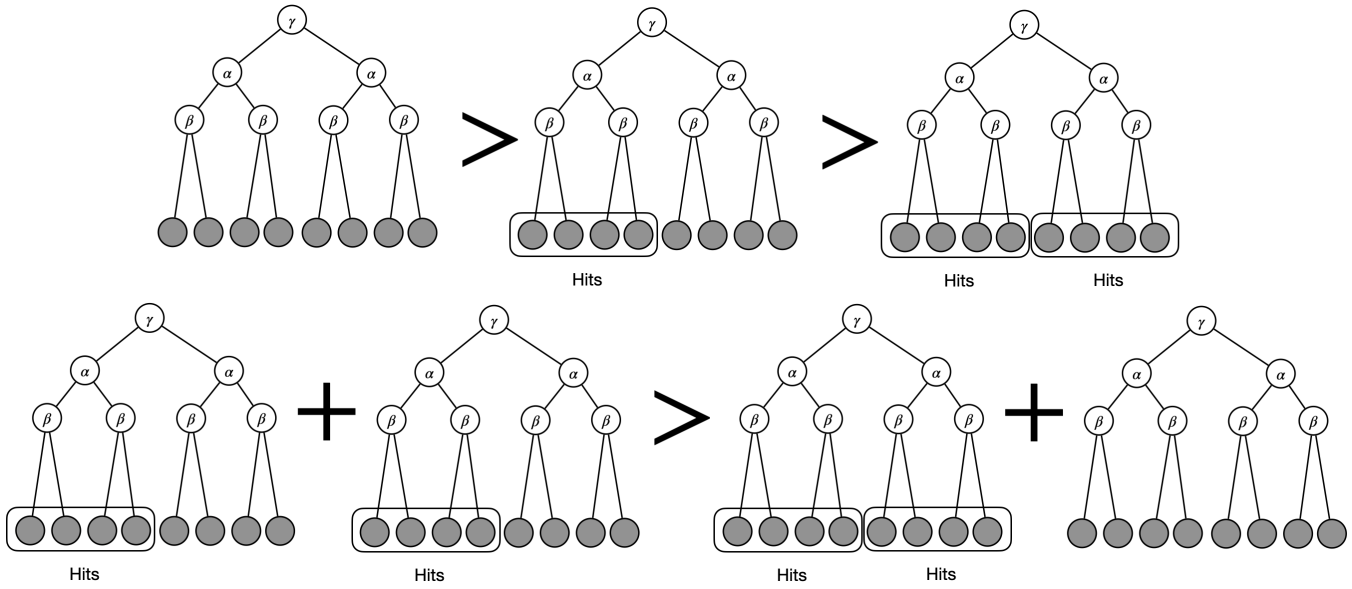


Figure 9: An illustration of three lemmas about 'hits'.

Lemma 4. *Let v and v' be two γ-type nodes. The left and right child of v are u and w. The left and right child of v' are u' and w'. Suppose that of $C(u)$, $C(w)$ and $C(w')$ do not hits E' and $C(u')$ hits E'. Then, we have*

$$L^*(v) + L^*(u) + L^*(w) > L^*(v') + L^*(u') + L^*(w')$$

Proof. We denote $|\mathcal{D}|$ as D. By the construction of $\mathcal{D}$ and $\mathcal{S}$, we have

$$D \geq 5, N = |\mathcal{S}| \geq D^3$$

By Claim 3 and Claim 1, we have

$$\begin{aligned}
&LHS - RHS \\
=&L^*(u) - L^*(u') - L^*(v') \\
=&(N-2D)\log\frac{N}{N-2D} + 2D\log\frac{N}{D} \\
&-((N+1-2D)\log\frac{N-1}{N+1-2D} + 2(D-1)\log\frac{N-1}{D-1}) \\
&-((N-1)\log\frac{N}{N-1} + \log N) \\
=&(N+1-2D)\log(N+1-2D) + 2(D-1)\log(D-1) \\
&-(N-2D)\log(N-2D) - 2D\log D
\end{aligned}$$

Let $\Delta L = \ln 2(LHS - RHS)$, we have

$$\frac{\mathrm{d}\Delta L}{\mathrm{d}N} = \ln(N+1-2D) - \ln(N-2D) > 0$$

Since $N \geq D^3$, we have

$$\begin{aligned}
&\ln 2(LHS - RHS) \\
\geq&(D^3+1-2D)\ln(D^3+1-2D) + 2(D-1)\ln(D-1) \\
&-(D^3-2D)\ln(D^3-2D) - 2D\ln D \\
\triangleq& f(D)
\end{aligned}$$

$$\begin{aligned}
&f'(D) \\
=&(3D^2-2)\ln(D^3-2D+1) - (3D^2-2)\ln(D^3-2D) \\
&+2\ln(D-1) - 2\ln D
\end{aligned}$$

$$\begin{aligned}
&f''(D) \\
=&6D\ln(D^3-2D+1) - 6D\ln(D^3-2D) \\
&+\frac{(3D^2-2)^2}{D^3-2D+1} - \frac{(3D^2-2)^2}{D^3-2D} + \frac{2}{D-1} - \frac{2}{D} \\
=&6D\ln\frac{D^3-2D+1}{D^3-2D} + \frac{(3D^2-2)^2}{D^3-2D+1} - \frac{(3D^2-2)^2}{D^3-2D} + \frac{2}{D-1} - \frac{2}{D} \\
\leq&6D(\frac{D^3-2D+1}{D^3-2D} - 1) + \frac{(3D^2-2)^2}{D^3-2D+1} - \frac{(3D^2-2)^2}{D^3-2D} + \frac{2}{D-1} - \frac{2}{D} \\
=&\frac{6}{D^2-2} - \frac{(3D^2-2)^2}{(D^3-2D+1)(D^3-2D)} + \frac{2}{(D-1)D} \\
=&-\frac{D^3-2D^2+6D-2}{(D-1)(D^2-2)(D^2+D-1)} \\
\leq&-\frac{2\cdot D^2-2D^2+6\cdot 2-2}{(D-1)(D^2-2)(D^2+D-1)} \\
=&-\frac{10}{(D-1)(D^2-2)(D^2+D-1)} \\
<&0
\end{aligned}$$

When $D \geq 5$, since $f''(D) < 0$, we have

$$\begin{aligned}
&f'(D)\\
&\geq \lim_{D\to\infty} f'(D)\\
&= \lim_{D\to\infty} [(3D^2-2)\ln(D^3-2D+1) - (3D^2-2)\ln(D^3-2D)\\
&\quad + 2\ln(D-1) - 2\ln D]\\
&\geq \lim_{D\to\infty} [2\ln(D-1) - 2\ln D]\\
&= \lim_{D\to\infty} [2\ln(1-\frac{1}{D-1})]\\
&=0
\end{aligned}$$

Since $f'(D) \geq 0$, we have

$$f(D) \geq f(5) > 0$$

and $LHS - RHS > 0$. ∎

Lemma 5. *Let v and v' be two γ-type nodes. The left and right child of v are u and w. The left and right child of v' are u' and w'. Suppose that of $C(u)$, $C(u')$ and $C(w')$ hit E' and $C(w)$ does not hits E'. Then, we have*

$$L^*(v) + L^*(u) + L^*(w) > L^*(v') + L^*(u') + L^*(w')$$

Proof. We denote $|\mathcal{D}|$ as D. By Claim 3 and Claim 1, we have

$$\begin{aligned}
&LHS - RHS\\
=&L^*(v) + L^*(w) - L^*(v') - L^*(w')\\
=&(N-1)\log\frac{N}{N-1} + \log N\\
&+ (N-2D)\log\frac{N}{N-2D} + 2D\log\frac{N}{D}\\
&- ((N-2)\log\frac{N}{N-2} + 2\log N)\\
&- ((N+1-2D)\log\frac{N-1}{N+1-2D} + 2(D-1)\log\frac{N-1}{D-1})\\
=&N\log N + (N-2)\log(N-2) - 2(N-1)\log(N-1)\\
&+ (N-2D+1)\log(N-2D+1) - (N-2D)\log(N-2D)\\
&+ 2(D-1)\log(D-1) - 2D\log D
\end{aligned}$$

In Lemma 3, we have seen that

$$N\log N + (N-2)\log(N-2) - 2(N-1)\log(N-1) > 0$$

In Lemma 4, we have seen that

$$\begin{aligned}
&N\log N + (N-2)\log(N-2) - 2(N-1)\log(N-1) > 0\\
&(N-2D+1)\log(N-2D+1) - (N-2D)\log(N-2D)\\
&+ 2(D-1)\log(D-1) - 2D\log D\\
>&0
\end{aligned}$$

Combining the above two inequalities, we have $LHS > RHS$. ∎

There are 2^l numbers of γ-type nodes and we denote them as $v_1, v_2, \ldots, v_{2^l}$. Let u_i and w_i be the left and right child of v_i. For an allocation of items to leaf nodes, we denote the number of hits as follows.

$$N_h = \sum_{i=1}^{2^l} \mathbb{I}[C(u_i) \text{ hits } E'] + \mathbb{I}[C(w_i) \text{ hits } E']$$

Let's consider the optimal $L^*(\mathcal{S})$ when N_h is fixed. Let

$$C_1 = (N+|\mathcal{D}|-2|\mathcal{F}|)\log\frac{N-|\mathcal{D}|}{N+|\mathcal{D}|-2|\mathcal{F}|} + 2(|\mathcal{F}|-|\mathcal{D}|)\log\frac{N-|\mathcal{D}|}{|\mathcal{F}|-|\mathcal{D}|}$$

, we have

$$\sum_{\beta\text{-type}v} L^*(v) = \sum_{\beta\text{-type}v} C_1 = 2^{l+1}C_1$$

By Lemma 3, when the smallest $L^*(S)$ obtained, there can be only one v_i such that only one among $C(u_i)$ and $C(w_i)$ hits E'.

Let

$$C_2 = (N+1-2|\mathcal{D}|)\log\frac{N-1}{N+1-2|\mathcal{D}|} + 2(|\mathcal{D}|-1)\log\frac{N-1}{|\mathcal{D}|-1}$$

$$C_3 = (N-2|\mathcal{D}|)\log\frac{N}{N-2|\mathcal{D}|} + 2|\mathcal{D}|\log\frac{N}{|\mathcal{D}|}$$

$$C_4 = (N-2)\log\frac{N}{N-2} + 2\log N$$

$$C_5 = (N-1)\log\frac{N}{N-1} + \log N$$

(1) When N_h is even.

$$L^*(\mathcal{S}; N_h) = 2^{l+1}C_1 + N_h C_2 + (2^{l+1} - N_h)C_3 + \frac{N_d}{2}C_4$$

(2) When N_h is odd.

$$L^*(\mathcal{S}; N_h) = 2^{l+1}C_1 + N_h C_2 + (2^{l+1} - N_h)C_3 + \frac{N_d - 1}{2}C_4 + C_5$$

By Lemma 4 and Lemma 5, we have $L^*(\mathcal{S}; N_h)$ increase with N_h. Finally, we set the input L of problem 3 as

$$L = L^*(\mathcal{S}; N_h = s)$$

If there exists a size-s matching M' of E', then we can construct an allocation such that $L^*(\mathcal{S}) \leq L$. When s is even, we set those item in hype-edge of M' to be $C(u_i)$ and $C(w_i)$, $i = 1, \ldots, s/2$. When s is odd, we set those item in hype-edge of M' to be $C(u_i)$ and $C(w_i)$, $i = 1, \ldots, (s-1)/2$ and $C(u_{(s+1)/2})$.

Suppose there exists an allocation such that $L^*(\mathcal{S}) \leq L$. If the number of hits N_h of this allocation is smaller than s, then we have $L^*(\mathcal{S}) < L^*(\mathcal{S}; N_h) = L$, which is contradict to the assumption.

We finish the proof of Theorem 3.

F Visualization on Size Distribution

A better-estimated distribution brings a better compression rate. Here, we visualize the distribution on $|S|$ to verify that our method is a better estimator.

The distribution of our model is just the $P(S; r)$ defined in section 5.1. For the distribution of single-item-based method, let $q_i = \Pr(i \in S)$, then the $Q(S) = \Pi_{i\in S} q_i \cdot \Pi_{j\in V-S}(1-q_j)$.

The exact size distribution of $P(S; r)$ and $Q(S)$ can be calculated by ordinary generating function efficiently.

F.1 Calculation of the size distribution

F.1.1 Ordinary Generating Function. We define the ordinary generating function(OGF) of a size distribution as follows.

$$F(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots$$

where the a_i is the $\Pr(|S| = i)$, for $i \in \mathbb{N}$.

F.1.2 Calculation of the Size Distribution of Single-item-based Model. For $i \in V$, let $q_i = Pr(i \in S)$. The probability of a $S \subseteq V$ for the single-item-based model is

$$Q(S) = \Pi_{i \in S} q_i \cdot \Pi_{j \in V-S}(1 - q_j)$$

We can calculate the OGF of $Q(S)$ by the following steps.

(1) F(x)=1
(2) For $i = 1, \ldots, K$,

$$F(x) := F(x) \times ((1 - q_i) + q_i x)$$

With $O(K)$ steps, we can obtain the size distribution of $Q(S)$.

F.1.3 Calculation of the Size Distribution of binary tree Model. For nodes l in the binary tree, we can recursively calculate the OGF $F_v(x)$ of $P(S; v)$. The notations are the same as section 5.1.

For a leaf node l, the OGF $F_l(x) = x$. For an internal node v, let v's left and right child be u and w, we have

$$F_v(x) = p(0; v)F_u(x) + p(1; v)F_w(x) + p(2; v)F_u(x)F_w(x)$$

With $O(2K + 1)$ steps, we can obtain the OGF $F_r(x)$ of the root node r.

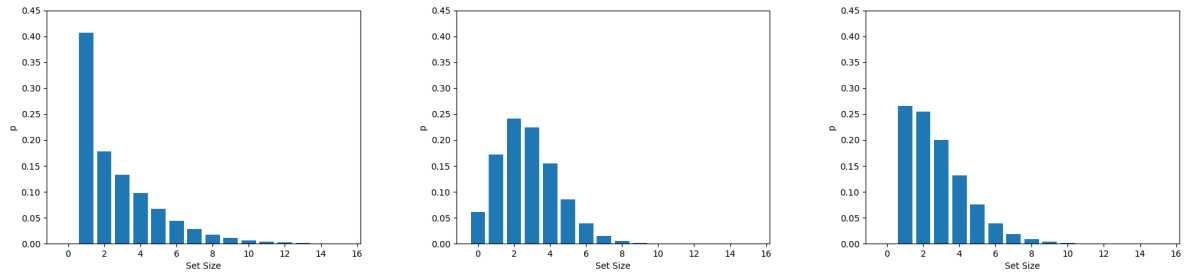
F.2 Results of Fitting Set Size Distribution

Fig. 10 and Fig. 11 visualize the size distribution of ground truth distribution $P^*(S)$, $Q(S)$ and $P(S; r)$ on the Tmall dataset. As illustrated in the figures, the $Q(S)$ has a positive value on $|S| = 0$, while there is zero probability for $|S| = 0$ in $P^*(S)$ and $P(S; r)$.

Both the ground truth size distribution and our model's size distribution have a similar fast-decaying shape. While the size distribution of the single-item-based model has a different bell shape like Gaussian.

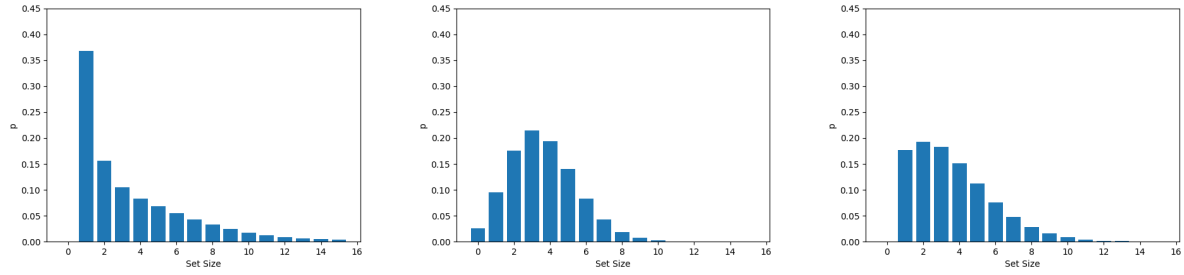
The visualization of the size distribution further confirmed that our binary tree model matches the pattern of real-world data better and thus brings us a lower compression rate.

Received 1 Feb 2025



(a) Size distribution in dataset $\mathcal{S}$ (b) Size distribution of single-item-based (c) Size distribution of Ours

Figure 10: The size distribution on the Tmall dataset. We only present the proportion of $|S| \leq 16$. Fig. (10a): The empirical size distribution in the Tmall dataset. Fig. (10b): The size distribution of the single-item-based baseline model learned by the Tmall dataset. Fig. (10c): The size distribution of the binary tree model learned by the Tmall dataset.



(a) Size distribution in dataset $\mathcal{S}$ (b) Size distribution of single-item-based (c) Size distribution of Ours

Figure 11: The size distribution on the HKTVmall dataset. We only present the proportion of $|S| \leq 16$. Fig. (11a): The empirical size distribution in the HKTVmall dataset. Fig. (11b): The size distribution of the single-item-based baseline model learned by the HKTVmall dataset. Fig. (11c): The size distribution of the binary tree model learned by the HKTVmall dataset.