arXiv:2302.02586v1 [cs.DS] 6 Feb 2023

最优 LZ-End 解析是困难的

HideoBannai[1] Mitsuru Funakoshi[2,3] Kazuhiro Kurita[4]Yuto Nakashima[2] Kazuhisa Seto[5] Takeaki Uno[6]

[1] 东京医科齿科大学 M&D 数据科学中心，日本东京 hdbn.dsc@tmd.ac.jp[2] 九州大学信息学系，日本福冈 mitsuru.funakoshi@inf.kyushu-u.ac.jp nakashima.yuto.003@m.kyushu-u.ac.jp[3] 日本学术振兴会，日本东京[4] 名古屋大学，日本名古屋，kurita@i.nagoya-u.ac.jp [5] 北海道大学信息科学与技术学院，日本札幌 seto@ist.hokudai.ac.jp[6] 日本国立信息学研究所，日本东京 uno@nii.jp

2023年2月7日

摘要

LZ-End 是著名的 Lempel-Ziv 解析系列的一种变体，其解析中的每个短语都有一个先前出现的位置，并且额外要求该先前出现的位置必须恰好结束于某个先前短语的末尾。LZ-End 最初被提出为一种贪心解析：从左到右贪心地确定每个短语，使其成为满足上述约束的最长因子 [Kreft 和 Navarro， 2010]。在本文中，我们研究短语数量最少的最优 LZ-End 解析。我们通过从顶点覆盖问题进行归约，证明计算最优 LZ-End 解析的判定版本是 NP 完全的。此外，我们借鉴近期提出的一种计算多种 NP 难重复性度量的方法，给出最优 LZ-End 解析的最大可满足性表述 [Bannai 等人， 2022]。我们还研究贪心 LZ-End 解析的规模与最优 LZ-End 解析的规模之比，并给出该比率的一个下界，该下界渐近趋近于 2。

1 引言

在无损数据压缩领域，各种重复性度量——尤其是基于字典压缩算法的度量——及其相互关系近来受到广泛关注（参见 Navarro 的精彩综述 [11, 12]）。最基本且最著名的度量之一是 LZ77 解析 [14], ，它将一个字符串解析为 z

若干短语，使得每个短语要么是单个符号，要么是此前出现过的最长子串。LZ-End [8, 9]是 LZ77 解析的一种变体，其额外限制是：短语此前出现的位置必须恰好结束于某个先前短语的末尾。更形式化地说，LZ-End 解析是给定字符串的一系列子串（称为短语） q_1，.....，q_{z_e}，可通过从左到右的贪心方式得到：每个短语 q_i 满足以下条件之一：(1) q_i 是一个符号，且该符号在字符串中首次出现于此处；或 (2) q_i 是剩余后缀的最长前缀，并且是某个 $j < i$的后缀，其中 $q_1 \cdots q_j$。已知 LZ-End 解析可在线性时间内计算 [6], ，且存在一种空间高效算法 [5]。

目前尚无已知大小为 $O(z)$、能够高效随机访问字符串中任意位置的数据结构~，但最近有研究表明，可以用 $O(z_e)$ 的空间实现 $O(1)$ 时间访问 [7]。此外，关于 z 与 z_e之间的差异，已有研究证明其上界为 $z_e = O(z\log^2(n/z))$ [7], ，其中 n 为（未压缩）字符串的长度。另一方面，对于一元字符串，显然有下界 $z_e = \Omega(z \log n)$，因为 LZ-End 短语的先前出现位置不能自引用，即不能与自身重叠，而 LZ77 短语则可以。注意，对于任意字符串，$z \leq z_{no} \leq z_e$ 均成立，其中 z_{no}是不允许自引用的 LZ77 解析中的短语数。已知存在一族字符串，使比率 z_e/z_{no} 渐近趋近于 2（对于较大的字母表 [9], ，二进制字母表 [4]），并且人们猜想 $z_e \leq 2z_{no}$ 对任意字符串 [9]均成立。

虽然上述解析中的短语是以贪心方式（即选择最长者）选取的，但我们也可以考虑不施加此类约束的变体。例如，在 类 *LZ-End* 解析中，每个短语 q_i 满足以下条件之一：（1） q_i 是一个单独的符号，且该符号是最左侧出现的符号；或（2） q_i 是剩余后缀的一个前缀（不一定是最长前缀），且是 $q_1 \cdots q_j$的某个后缀，其中 $j < i$。我们将短语数量最少的 LZ-End-like 解析称为 z_{end} 最优 LZ-End 解析 [11], ，并将原始解析称为贪心 LZ-End 解析。[1] 因此，$z \leq z_{no} \leq z_{end} \leq z_e$ 成立。有趣的是， $z_{end} \leq g$ 也成立，其中 g 是能够（仅）生成该字符串的最小上下文无关文法的大小，而 z_e 与g 之间类似的关系似乎尚未为人所知 [11]。

这引出了关于度量 z_{end}的两个自然而重要的问题：

- 我们能以多高的效率计算 z_{end}?
- z_{end} 相比 z_e最多能缩小多少？本文回答了上述问题中的一部分，具体如下：1. 我们证明了计算 z_{end}的 NP 困难性。2. 我们提出了一种通过最大可满足性进行精确计算的算法。3. 我们给出了比率 z_e/z_{end}最大值的下界。

在第3节中，我们给出困难性结果。我们的归约来自顶点覆盖问题：寻找一个最小的顶点集合 U ，使得每条边都与 U 中的某个顶点关联。在第4节中，我们给出一种用于计算最优 LZ-End 解析的最大可满足性表述，该表述采用了 Bannai 等人的方法，能够借助最大可满足性求解器计算 NP 难重复性度量 [1]。在第5节中，我们考察比率 z_e/z_{end}。我们给出一个二进制字符串族，使得该比率渐近趋近于 2。请注意，我们可以轻松地将这一结果推广到更大的字母表。由于 $(z_e/z_{end}) \leq (z_e/z_{no})$，假设 Kreft 和 Navarro [9] 的猜想成立，则该界是紧的，

[1]请注意，对于 LZ77，我们无需作此区分，因为贪心 LZ77 解析也是一种最优 LZ77 类解析。

相关工作

LZ77 和 LZ78 是 LZ 家族 [14, 15] 中最初的成员。众所周知，（贪心）LZ77 解析会产生该解析的最优版本 [10]。另一方面，计算 LZ78 变体 [2] 的最优版本是 NP 困难的。这一困难性结果同样通过从顶点覆盖问题归约得到。然而，我们针对 LZ-End 构造的归约与针对 LZ78 的归约不同，因为这两种解析的结构迥异。

2 预备知识

2.1 字符串

设 Σ 为一个字母表。Σ^* 中的一个元素称为字符串。字符串 w 的长度记为 $|w|$。空字符串 ε 是长度为 0 的字符串。设 Σ^+ 为非空字符串的集合，即 $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$。对于任意字符串 x 和 y，$x \cdot y$ 表示两个字符串的连接。我们有时将“$\cdot$”（即 $x \cdot y = xy$）简写。对于字符串 $w = xyz$，x、y 和 z 分别称为 w的前缀、子串和后缀。如果 $x \neq w$、$y \neq w$和 $z \neq w$ 分别成立，则它们称为 w 的真前缀、真子串和真后缀。字符串 w 的第 i个符号记为 $w[i]$，其中 $1 \leq i \leq |w|$。对于字符串 w 和两个整数$1 \leq i \leq j \leq |w|$，令 $w[i..j]$ 表示 w 中从位置 i 开始到位置 j 结束的子串。为方便起见，当 $i > j$ 时，令 $w[i..j] = \varepsilon$ 。我们有时用 $w[i..j)$ 表示 $w[i..j-1]$。对于任意字符串 w，令 $w^1 = w$ ，并对任意整数 $k \geq 2$，令 $w^k = ww^{k-1}$ ，即 w^k 是 w重复 k次所得的字符串。

2.2 LZ-End 解析

We g给出 LZ-End 解析的定义，它是 Lempel-Ziv 家族的一种变体。

定义 1 （（贪心）LZ-End 解析）。 字符串 w 的LZ-End 解析 是解析 $\mathit{LZEnd}(w) = q_1, \ldots, q_{z_{end}}$，其中 q_i 要么是符号最左侧出现位置上的单个符号，要么是 $q_i \cdots q_{z_{end}}$ 的最长前缀，且该前缀作为后缀出现在某个 $q_1 \cdots q_j$ 中；我们称这样的 $j < i$为该短语的来源 。

此定义用于 [7], ，与原始版本 [8, 9] 略有不同，原始版本会在每个短语中添加一个符号。本文的结果同样适用于原始版本（我们将在论文完整版中证明这一点），但为简洁起见，此处采用此定义。我们将每个 q_i 称为一个短语。本文考虑 LZ-End 解析的一个更一般版本：字符串 w的解析 $q_1, \ldots, q_{z_e}$，其中 q_i 是某个 $q_1 \cdots q_j$ 的一个（不一定是最长）后缀。我们将短语数最少的此类解析 z_e 称为 w的最优 *LZ-End* 解析。图1给出了贪心 LZ-End 解析和最优 LZ-End 解析的示例。

2.3 图

设 $G = (V, E)$ 是一个图，其顶点集为 V，边集为 E。边 $e = \{u, v\}$ 称为 u 的关联边。我们将 v 的关联边集合记为 $\Gamma_G(v)$。若不会引起混淆，我们会省略下标。对于边 $e = \{u, v\}$，顶点 u 和 v 是 e 的端点。对于顶点子集 $U \subseteq V$，若对于任意

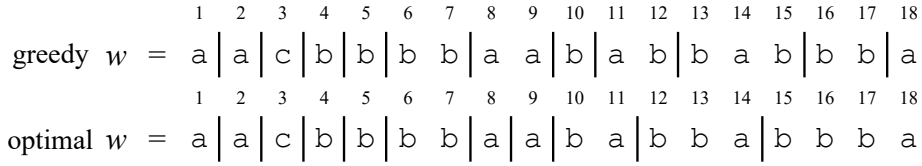

图1：设 $w =$ 为 aacbbbbaababbabbba。图的上半部分展示了 w 的贪心 LZ-End 解析 $LZEnd(w)$。对于位置 10 处的短语，更长的子串 $w[10..11] =$ ba 在位置 7 还有一次先前出现，但没有短语在位置 8 结束，而且任何更长的子串都没有先前出现。因此，位置 10 处的短语是 b。图的下半部分展示了同一字符串的一个最优 LZ-End 解析（其短语数少于贪心解析）。每个短语都有一次先前出现，且该出现结束于某个 LZ-End 短语的末尾。贪心解析的大小为 12，最优解析的大小为 11。

$e \in E$， e 的至少一个端点包含在 U中。设 τ_G 为 G 的最小顶点覆盖的大小（即， τ_G 表示 G的顶点覆盖数 ）。注意，计算 τ_G是 NP 完全的 [3]。

3 计算最优 LZ-End 解析的 NP 困难性 g

本节中，我们考虑计算给定字符串的最优 LZ-End 解析这一问题。该问题的判定版本如下。

问题2 (计算最优 LZ-End 解析的判定版本 (OptLE))。 给定一个字符串 w 和一个整数 k，判断是否存在大小不超过 k 的 *LZ-End* 解析。

下面我们证明 OptLE 的 NP 完全性，并在下一节给出一个用于精确计算的算法。

定理3。 OptLE是 *NP* 完全的。

证明。 我们将“顶点覆盖问题”归约到 OptLE。令 $G = (V, E)$ 为一个图，其中顶点集为 $V = \{v_1, \ldots, v_n\}$，边集为 $E = \{e_1, \ldots, e_m\}$。假设顶点覆盖问题的输入图 G 是连通的，并且对于任意$v \in V$， $|\Gamma(v)| \geq 2$ 。我们将每个顶点 v_i 视为一个符号 v_i，并将每条边 e_i 视为一个符号 e_i。我们还引入符号 \$，以及一组在字符串中各自只出现一次的符号。为简便起见，后者用特殊符号 # 表示；也就是说，在我们的描述中，# 每次出现都代表一个不同的符号。我们考虑由图 G 定义的字符串 $\mathcal{W}_G$，如下所示。

- $\mathcal{W}_G = \prod_{i=1}^{n} \mathcal{P}_i \cdot \prod_{j=1}^{m} \mathcal{Q}_j \cdot \prod_{i=1}^{n} \mathcal{R}_i \cdot \prod_{j=1}^{m} \mathcal{S}_j$
- $\mathcal{P}_i = v_i^3 \# v_i^2 \$ \# v_i \$^2 \# \cdot \mathcal{X}_i \cdot \mathcal{Y}_i$
- $\mathcal{Q}_j = e_j^3 \#$
- $\mathcal{R}_i = v_i^4 \$ \prod_{e_j \in \Gamma(v_i)} (e_j^3 v_i^2 \$) \$ \#$

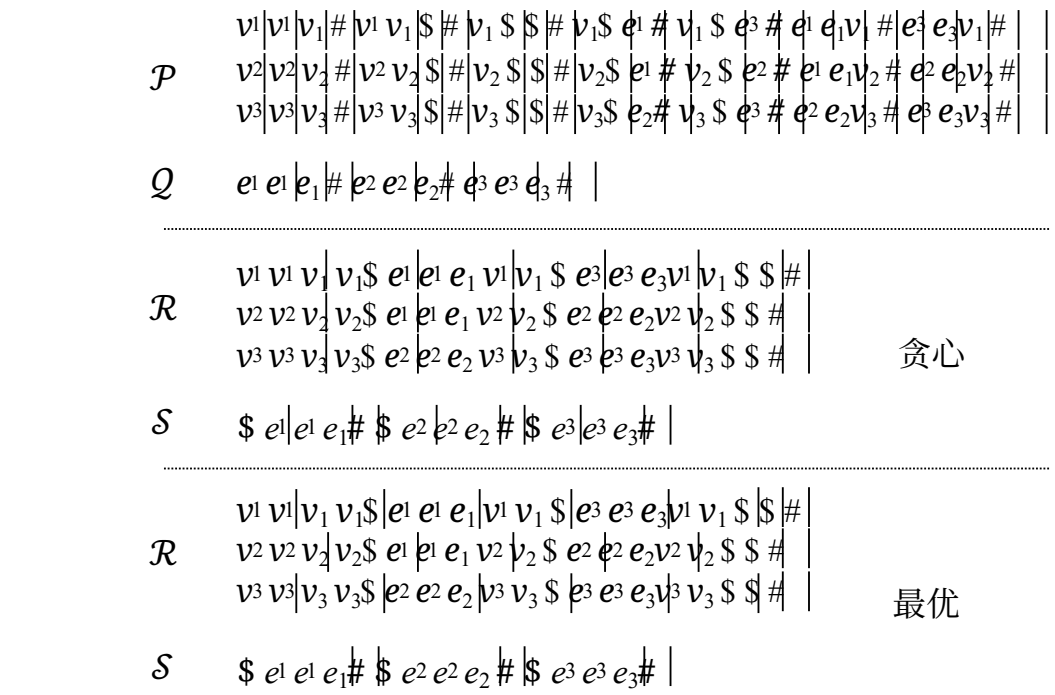

图2：令 $G=(V,E)$ 为一个含有三个顶点 v_1, v_2, v_3 的完全图，其顶点为 $e_1=\{v_1,v_2\}$, $e_2=\{v_2,v_3\}, e_3=\{v_1,v_3\}$。图中展示了 $\mathcal{W}_G$、贪心解析和一种最优解析。前两部分（$\mathcal{P}$ 和 $\mathcal{Q}$）采用相同的解析，后两部分（$\mathcal{R}$ 和 $\mathcal{S}$）则有所不同。图的上半部分展示贪心解析，下半部分展示一种最优解析。例如，在最优解析中，我们可以在 $\mathcal{R}_1$ 中采用非贪心解析，将 $\$e_1^3$ 和 $\$e_3^3$ 选作短语。换言之，我们可以通过在 $\mathcal{R}_1$ 中增加一个短语，减少 $\mathcal{S}$ 部分中的两个短语。在这个例子中，最优解析表示图 G 的一个顶点覆盖 $\{v_1,v_3\}\subset V$（因为 $\mathcal{R}_2$ 选择了贪心解析，而其余部分没有）。

- $\mathcal{S}_j = \$e_j^3\#$
- $\mathcal{X}_i = \prod_{e_j\in\Gamma(v_i)}(v_i\$e_j\#)$
- $\mathcal{Y}_i = \prod_{e_j\in\Gamma(v_i)}(e_j^2 v_i\#)$

该字符串的一个示例如图2所示。请注意，我们使用 i 表示顶点的索引，使用 j 表示边的索引。我们证明，当且仅当顶点覆盖数 τ_G 小于 k时， $\mathcal{W}_G$ 的最优解析中的短语数小于 $13n+22m+k$。

首先，我们观察 $\mathcal{W}_G$ 的一种最优 LZ-End 解析。考虑 $\prod_{i=1}^n \mathcal{P}_i$ 的一种解析。在这一部分，贪心解析得到 $10n+13m$ 个短语。在对 $\prod_{i=1}^n \mathcal{P}_i$ 进行贪心解析时， $v_i^2\$\#$ 中的短语 v_i^2、 $v_i\$^2\#$ $v_i\$e_j\#$ 中的短语 $v_i\$$，以及 e_j^2 的第二次出现的长度均为 2，其他短语的长度均为 1。显然，这种解析是对 $\prod_{i=1}^n \mathcal{P}_i$ 的最短解析。此外，其他规模相同的解析不会影响字符串其余部分的解析；选择其他解析也无法增加来源的候选项

因为长度为 2 的短语前面都有唯一符号 #。因此，我们可以选择这种贪心解析，作为某个最优解析的一部分。

在第二部分 $\prod_{j=1}^{m} \mathcal{Q}_j$ 中，贪心解析也给出一个最优解析，其中包含 $3m$ 个短语（即，每个 $\mathcal{Q}_j$ 都解析为三个短语，因为对于某个 i，e_j^2 会出现在 $\mathcal{P}_i$ 中，而 e_j^3 在 $\prod_{i=1}^{n} \mathcal{P}_i \cdot \prod_{j=1}^{m} \mathcal{Q}_j$ 中是唯一的）。这个解析也是可能的最小解析，并且不会影响字符串其余部分的任何解析。

剩余后缀 $\prod_{i=1}^{n} \mathcal{R}_i \cdot \prod_{j=1}^{m} \mathcal{S}_j$ 是归约的关键。关键思想是，$\mathcal{S}_j$ 表示边 e_j 是否是顶点子集中的某个顶点的关联边。$\mathcal{S}_j$ 中的 $\$e_j^3$ 在 $\mathcal{R}$ 部分中恰好有两个先前出现位置（因为每条边恰好关联两个顶点）。因此，当且仅当 $\$e_j^3$ 在 $\mathcal{R}$ 部分中有一个以 LZ-End 短语结尾的出现位置时，$\mathcal{S}_j$ 才能解析为两个短语（即，$\$e_j^3$, #）。现在我们考虑 $\mathcal{R}_i$ 部分的贪心解析（令 $\Gamma(v_i) = \{e_{(i,1)}, \ldots, e_{(i,|\Gamma(v_i)|)}\}$），如下所示：

$$v_i^3, v_i\$e_{(i,1)}, e_{(i,1)}^2 v_i, \ldots, v_i\$e_{(i,|\Gamma(v_i)|)}, e_{(i,|\Gamma(v_i)|)}^2 v_i, v_i\$^2, \#.$$

该解析包含 $2|\Gamma(v_i)| + 3$ 个短语。我们断言，这种解析是可能的最小解析：如果每个短语的长度至多为 3，那么由于 $\mathcal{R}_i$ 的长度为 6（$|\Gamma(v_i)| + 1$）$+1$，因此 $2|\Gamma(v_i)| + 3$ 是最小规模。另一方面，根据定义，我们可以看到，整个字符串 $\mathcal{W}_G$ 中，包含符号 v_i 且长度至少为 4 的子串都是唯一的。也就是说，$\$e_j^3$ 是唯一一个长度至少为 4 且不唯一的子串。然而，任何在第二次出现时选择长度为 4 的子串的解析，都至少包含 $2|\Gamma(v_i)| + 4$ 个短语，因为这些长度为 4 的子串并未严格包含贪心解析中的任何短语。因此，$\mathcal{R}_i$ 的最小短语数为 $2|\Gamma(v_i)| + 3$，上述贪心解析是唯一可能达到最小规模的解析。请注意，这种解析中的短语都不以 $\$e_j^3$ 结尾。下面来考虑 $\mathcal{R}_i$ 部分的另一种可能解析：

$$v_i^2, v_i^2\$, e_{(i,1)}^3, \ldots, v_i^2\$, e_{(i,|\Gamma(v_i)|)}^3, v_i^2\$, \$, \#.$$

该解析包含 $2|\Gamma(v_i)| + 4$ 个短语。注意，该解析中有些短语以 $\$e_j^3$ 结尾。因此，如果我们选择一种非贪心解析，使得存在一个短语在这些位置之一结束，那么 $\mathcal{S}_j$ 就可以解析为两个短语。换句话说，如果我们在 $\mathcal{R}_i$ 部分选择这样的解析，那么在 $\mathcal{S}$ 部分至多可以减少 $|\Gamma(v_i)|$ 个短语。这些观察结果意味着，在 $\mathcal{W}_G$ 的任意最优解析中，$\mathcal{R}_i$ 会被解析为 $2|\Gamma(v_i)| + 3$ 个或 $2|\Gamma(v_i)| + 4$ 个短语。

考虑一个最优 LZ-End 解析。令 r 表示包含 $2|\Gamma(v_i)| + 4$ 个短语的子串 $\mathcal{R}_i$ 的数量，令 s 表示恰好包含两个短语的子串 $\mathcal{S}_j$ 的数量。那么该解析的大小为

$$(10n + 13m) + (3m) + (2\sum_{i=1}^{n} |\Gamma(v_i)| + 3n + r) + (3m - s) = 13n + 23m + r - s.$$

我们考虑一个顶点子集 V'，使得 $v_i \in V'$ 当且仅当 $\mathcal{R}_i$ 被解析为 $2|\Gamma(v_i)|+4$ 个短语（即 $|V'| = r$）；并考虑一个边子集 E'，使得 $e_j \in E'$ 当且仅当 $\mathcal{S}_i$ 被解析为三个短语（即 $|E'| = m - s$）。如果 $E' = \emptyset$（即 $s = m$），那么 V' 是 G 的顶点覆盖。否则，V' 不是 G 的顶点覆盖。不过，我们可以通过解析求得顶点覆盖数。由于该解析是最优解析，我们可以观察到，在 $V \setminus V'$ 中不存在这样的顶点 v_i：它在 E' 中有两条或更多条关联边（通过在 $\mathcal{R}_i$ 中添加一个短语，可以减少 $\mathcal{S}$ 部分中的两个或更多短语，这与最优性矛盾）。这意味着，我们可以为 $E \setminus E'$ 中的每条边从 $V \setminus V'$ 中选取一个顶点，从而得到一个顶点覆盖。于是

存在一个大小相同的最优 LZ-End 解析，可以直接表示一个顶点覆盖。换言之，如果存在一个包含 $13n + 22m +（r + m - s）$个短语的最优 LZ-End 解析，那么顶点覆盖数就是 $r+m-s$。显然，从上述构造可知，如果顶点覆盖数为 k，则存在一个包含 $13n + 22m + k$ 个短语的最优 LZ-End 解析。 □

4 最大可满足性表述

文献 [1] 展示了一种通过将各种 NP 难的重复性度量表述为最大可满足性实例，从而利用高效求解器进行精确计算的方法。本文将展示，该方法也可用于计算最优 LZ-End 解析。

设输入字符串为 $T[1..n]$，对于任意 $i \in [2, n]$，令 $M_i = \{j \mid 1 \leq j < i, T[j] = T[i]\}$。下文中，我们用 1 表示真，用 0 表示假。我们引入以下布尔变量：

- p_i 对于所有 $i \in [1, n]$: $p_i = 1$ 当且仅当位置 i 是某个 LZ-End 短语的起始位置。注意，$p_1 = 1$。
- c_i 对于所有 $i \in [1, n]$: $c_i = 1$ 当且仅当位置 i 是符号 $T[i]$ 的最左出现位置。
- $r_{i \to j}$ 对所有 $i \in [2, n]$ 和 $j \in M_i$: $r_{i \to j} = 1$ 当且仅当位置 i 通过 LZ-End 因子引用位置 j时成立。

请注意，对于给定的字符串，c_i 的真值都是固定的，并且很容易确定。此外，最左出现位置必须是一个短语的起始位置，因此，p_i的一些取值也可以固定。对于所有 $i \in [1, n]$:

$$c_i = p_i = 1 \text{ if } i \text{ is left-most occurrence of } T[i], \tag{1}$$
$$c_i = 0 \text{ otherwise.} \tag{2}$$

最大可满足性是可满足性问题的一种变体，其中包含两类子句：硬子句和软子句。最大可满足性问题的解是对变量进行真值赋值，使所有硬子句均得到满足（求值为 1），并使得到满足的软子句数量最大化。

p_i 的真值定义了因子，因此，为了使因子数量最少，我们将软子句定义为 $\neg p_i$，对所有 $i \in [1, n]$均成立。下面给出变量之间必须满足的其他约束，即硬子句。

任意位置上的符号必须是最左出现位置，或者引用其左侧的某个位置。也就是说，对于任意 $i \in [1, n]$:

$$c_i + \sum_{j \in M_i} r_{i \to j} = 1. \tag{3}$$

为了确保同一个 LZ-End 短语中的引用保持一致，我们有以下两个约束。第一个约束确保，如果 i 引用 j，且位置

$i-1$ 和 $j-1$ 处的符号不同或不存在（即 $j=1$），则位置 i 和 $i-1$ 不能处于同一个 LZ-End 短语中。对于所有满足 $j=1$ 或 $T[j-1]\neq T[i-1]$的 $i\in[2,n]$ 和 $j\in M_i$：

$$r_{i\to j} \implies p_i, \tag{4}$$

第二个约束确保：如果位置 i 引用位置 j，且 i 不是 LZ-End 短语的起始位置，那么位置 $i-1$ 必须引用位置 $j-1$。对于所有满足 $T[j-1]=T[i-1]$的 $i\in[2,n]$ 和 $j\in M_i\setminus\{1\}$：

$$r_{i\to j}\wedge\neg p_i \implies r_{i-1\to j-1}. \tag{5}$$

最后，以下约束确保每个 LZ-End 短语的引用必须结束于某个先前 LZ-End 短语的末尾。对于所有 $i\in[1,n]$ 和 $j\in M_i$：

$$\begin{cases} r_{i\to j}\wedge p_{i+1} \implies p_{j+1} & \text{if } i\in[1,n)\\ r_{i\to j} \implies p_{j+1} & \text{if } i=n. \end{cases} \tag{6}$$

显然，从任意 LZ-End 解析导出的真值赋值都会满足上述约束。

现在我们证明，任何满足上述约束的真值赋值都表示一个有效的 LZ-End 解析。p_i 的真值对应一种解析，其中每个短语都从位置 i 开始，当且仅当 $p_i=1$。约束 (3) 确保每个位置要么是单字符短语，要么引用一个唯一的先前位置。因此，接下来只需证明，给定因子中每个位置的引用方式是一致的（相邻位置引用相邻位置），并且以某个先前短语的末尾结束。

对于任意满足 $c_i=0$ 的位置 i，令 $j\in M_i$ 为满足 $r_{i\to j}=1$ 的唯一值。可以看出，任何不在短语开头（即 $p_i=0$）的此类位置 i，都会引用与位置 $i-1$ 的引用一致的位置：如果 $j=1$ 或 $T[j-1]\neq T[i-1]$，那么约束 (4) 将推出 $r_{i\to j}=0$。因此，我们有 $j>1$ 和 $T[j-1]=T[i-1]$；根据约束 (5)，可得 $r_{i-1\to j-1}$，并且因子内部的引用是一致的。最后，根据约束 (6)，短语中的最后一个引用始终指向某个先前 LZ-End 短语的末尾。

该 MAX-SAT 实例包含 $O(n^2)$ 个变量，合取范式的总规模为 $O(n^2)$：$O(n)$ 个规模为 $O(n)$ 的子句（约束 (3) 使用了基数约束的线性规模编码，例如 [13]），以及 $O(n^2)$ 个规模为 $O(1)$ 的子句（软子句以及约束 (4)、(5)、(6)）。

我们注意到，只需稍作修改，就不难为 LZ-End 的原始定义构造一个最大可满足性表述。

5 贪心解析与最优解析的近似比

本节中，我们考察贪心 LZ-End 解析的大小 z_e 与最优 LZ-End 解析的大小 z_{end} 之比。这里，我们给出该比率的一个下界。

定理 4. 存在一个二进制字符串族，使得该比率 z_e/z_{end} 渐近趋近于 2。

证明。令 $K=\sum_{i=1}^{k}2^i$ 对于任意正整数 $k\geq 1$。字母表 $\{a,b\}$ 上的以下二进制字符串 w_k 给出了下界：

$$w_k = aa\cdot\prod_{i=1}^{k}(a^{2^i})\cdot b^4\cdot\prod_{i=1}^{K}(a^ib^3).$$

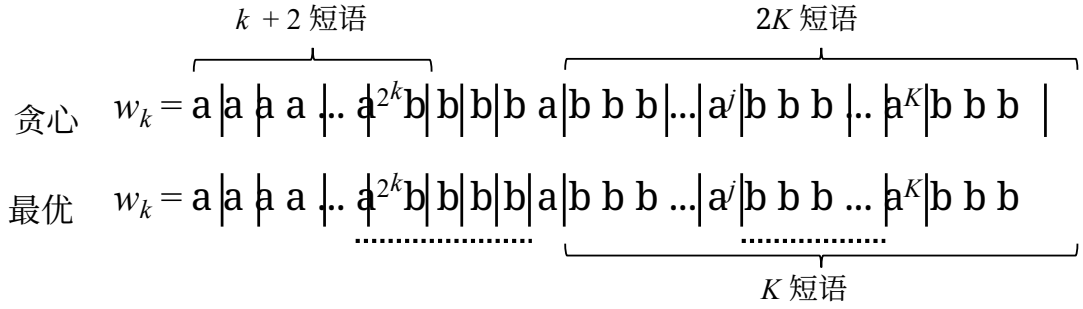

图 3：字符串 w_k 的两种 LZ-End 解析变体示意图（定理 4）。在最优解析中，对于每个 j $(1 \le j \le K)$，我们可以选择 $a^j b^3$ （虚线）作为一个短语，只需添加一个单字母短语 b。

容易看出， K 是子串 $\prod_{i=1}^{k}(a^{2^i})$ 的长度。首先，我们展示 w_k的贪心解析。令 $W_0 = aa \cdot \prod_{i=1}^{k}(a^{2^i}) \cdot b^4$ （即 w_k的前缀），并令 $W_j = W_{j-1} \cdot a^j b^3$ 对于任意$1 \le j \le K$。注意到 $W_K = w_k$。我们证明

$$LZEnd(W_j) = LZEnd(W_{j-1}), a^j b^2, b \tag{7}$$

根据 j 进行归纳。首先，考虑 W_0 的贪心解析。第一次游程（即包含唯一符号的极大子串）的贪心解析为 a， a， a^2，……， a^{2^k}，大小为 $k+2$。第二次游程被解析为三个短语 b， b， b^2。因此

$$LZEnd(W_0) = a, a, a^2, \ldots, a^{2^k}, b, b, b^2.$$

此外，LZEnd (W_1) =LZEnd (W_0) ab^2, b，因为 $W_1 = W_0 \cdot ab^3$ 成立。因此，等式 7 对 $j = 1$ 成立。假设对于某个整数 $p \ge 1$，等式 7 对任意 $j \le p$ 都成立。我们证明等式 7 对 $j = p+1$ 也成立。假设 LZEnd(W_{p+1}) 中存在一个短语 x，它始于 W_p，止于 W_{p+1}的一个新后缀 $a^{p+1}b^3$ 。根据归纳假设，LZEnd (W_p) 中以 a 结尾的短语仅位于第一次 a 游程中。这意味着 x 不能以 a 结尾，并且 x 可以写成 $x = x'ba^{p+1}b^{\ell}$，其中 x' 是 x 的一个前缀， ℓ 是某个正整数。然而， a^{p+1} 只出现在第一次 a 游程中。因此，LZEnd (W_{p+1}) 不可能包含这样的短语 x；也就是说，对于 S 的某种因式分解，LZEnd (W_{p+1}) =LZEnd(W_p) S。容易看出， W_{p+1} 的剩余后缀 $a^{p+1}b^3$ 被解析为 $a^{p+1}b^2, b$。因此，等式 7 对 $j = p+1$ 成立，对任意 j 也成立。注意， |LZEnd(w_k)| $= 2K + k + 5$ 成立。

最后，我们给出一个更小的 w_k解析。对于第一次运行，我们采用相同的解析；对于第二次运行，我们采用不同的解析，如 b， b， b， b。在贪心解析中， $a^j b^3$ 不能是一个短语，因为它唯一的先前出现位置并不以 LZ-End 短语结尾。我们可以使用子串 $a^j b^3$ 作为 W_j 中的一个新短语（另见图 3）。因此存在一个 LZ-End 解析

$$a, a, a^2, \ldots, a^{2^k}, b, b, b, b, a^1 b^3, \ldots, a^K b^3.$$

该解析的大小为 $K + k + 6$。

因此，对于这个字符串族，比率 z_e/z_{end} 渐近趋近于 2。 □

注意，这个字符串族也给出了比率 z_e/z_{no} 的下界，因为 (z_e/z_{end}) ≤(z_e/z_{no}) 成立。

6 结论

本文首先研究了 LZ-End 变体的最优版本。我们证明，计算最优 LZ-End 解析的判定版本是 NP 完全的，并提出了一种将问题表述为最大可满足性实例、从而精确计算最优 LZ-End 的方法。我们还给出了贪心 LZ-End 与最优 LZ-End 之间可能差距（以比率表示）的下界。最后，我们指出了以下可能的未来研究方向。

- 我们从顶点覆盖问题出发的归约使用了多项式规模字母表。如何构造使用小字母表的归约?

- 最有意思的未解问题是第5节中讨论的比率的上界。我们猜想存在一个常数上界（即，$z_e/z_{end} \leq c$对于任何满足 c 为常数的字符串）。这意味着贪心解析能够对最优解析实现常数近似。另一方面，如果存在一个字符串族，使得 $c > 2$ 或比率不是常数，那么猜想 $z_e \leq 2z_{no}$就不成立。

致谢

感谢 Dominik K¨oppl 与我们讨论。

引用

[1] H. Bannai、K. Goto、M. Ishihata、S. Kanda、D. K¨oppl 和 T. Nishimoto。通过最大可满足性计算 NP-hard 重复性度量。载于 S. Chechik、G. Navarro、E. Rotenberg 和 G. Herman 编，第 *30* 届欧洲算法年会，*ESA 2022*，*2022* 年 *9* 月 *5* 日至 *9* 日，德国柏林/波茨坦，第244卷，*LIPIcs*，第12:1–12:16页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2022年。[2] S. De Agostino 和 J. A. Storer。动态文本压缩中的在线计算与离线计算。《信息处理快报》，59(3):169–174，1996年。[3] M. R. Garey 和 D. S. Johnson。《计算机与难解性》，第174卷。Freeman出版社，旧金山，1979年。[4] T. Ideue、T. Mieno、M. Funakoshi、Y. Nakashima、S. Inenaga 和 M. Takeda。关于 LZ-End 相对于 LZ77 的近似比。载于 T. Lecroq 和 H. Touzet 编，字符串处理与信息检索 —— 第 *28* 届国际研讨会，*SPIRE 2021*，法国里尔，*2021*年*10*月*4*日至*6*日，论文集，第12944卷，《计算机科学讲义》，第114–126页。施普林格，2021年。[5] D. Kempa 和 D. Kosolobov。压缩空间中的 LZ-End 解析。载于 A. Bilgin、M. W. Marcellin、J. Serra-Sagrist`a 和 J. A. Storer 编，*2017*年数据压缩会议，*DCC 2017*，斯诺伯德，犹他州，美国，*2017*年*4*月*4*日至*7*日，第350–359页。电气电子工程师学会，2017年。[6] D. Kempa 和 D. Kosolobov。线性时间的 LZ-End 解析。载于 K. Pruhs 和 C. Sohler 编，第*25*届欧洲算法年会，*ESA 2017*，*2017*年*9*月*4*日至*6*日，奥地利维也纳，第87卷，*LIPIcs*，第53:1–53:14页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2017年。[7] D. Kempa 和 B. Saha。LZ-End 解析的一个上界及线性空间查询。载于 J. S. Naor 和 N. Buchbinder 编，*2022*年 *ACM-SIAM* 离散算法研讨会论文集，*SODA 2022*，线上会议 / 美国弗吉尼亚州亚历山德里亚，*2022*年*1*月*9*日至*12*日，第2847–2866页。工业与应用数学学会，2022年。[8] S. Kreft 和 G. Navarro。支持快速随机访问的类LZ77压缩。载于 J. A. Storer 和 M. W. Marcellin 编，*2010*年数据压缩会议（*DCC 2010*），*2010*年*3*月*24*日至*26*日，美国犹他州斯诺伯德，第239–248页。IEEE计算机学会，2010年。[9] S. Kreft 和 G. Navarro。关于重复序列的压缩与索引。《理论计算机科学》，483:115–133，2013年。[10] A. Lempel 和 J. Ziv。关于有限序列的复杂度。《IEEE信息论汇刊》，22(1):75–81，1976年。[11] G. Navarro。高度重复字符串集合的索引。CoRR，abs/2004.02781。[12] G. Navarro。高度重复字符串集合的索引，第一部分：重复性度量。《ACM计算调查》，54(2):29:1–29:31，2021年。

[13] C. Sinz. 关于布尔基数约束的最优合取范式编码。载于 P. van Beek 编，约束编程原理与实践——*CP 2005*，第*11*届国际会议，*CP 2005*，西班牙锡切斯，*2005*年*10*月*1*日至*5*日，论文集，《计算机科学讲义》第3709卷，第827–831页。施普林格，2005年。[14] J. Ziv 和 A. Lempel. 一种用于序列数据压缩的通用算法。*IEEE*信息论汇刊，23(3):337–343，1977年。[15] J. Ziv 和 A. Lempel. 通过变速率编码压缩单个序列。*IEEE*信息论汇刊，24(5):530–536，1978年。