RC23254 (W0406-127) June 21, 2004
Electrical Engineering

IBM Research Report

Bounds on Expansion in LZ'77-Like Coding

Vittorio Castelli, Luis Alfonso Lastras-Montaño
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Bounds on Expansion in LZ'77-like coding

Vittorio Castelli and Luis Alfonso Lastras-Montaño

Abstract

We investigate the maximum increase in number of phrases that results from changing k consecutive symbols in a string $\mathbf{x}$ having length n parsed using an LZ'77-like algorithm. We consider a class of compression algorithms that partition a sequence into a collection $\mathbf{y}$ of non-overlapping, variable-length phrases and encode them. Each phrase is either a singleton or matches a substring that starts to its left.

We show that changing a single symbol of $\mathbf{x}$ in position i can yield an expansion that is of order $O(n-i)^{2/3}$ as $(n-i) \to \infty$. Our lower bound requires an alphabet size of $O(n-i)^{1/3}$. We also show that changing k consecutive symbols starting from position i can yield a expansion having a similar but somewhat more involved form. The paper contains both analytically-derived upper and lower bounds, and algorithms for numerically computing tighter bounds. While deriving the bounds, we provide a detailed analysis of how expansion can arise when changing consecutive symbols.

This problem is motivated by management policies for computer systems, such as the IBM Memory eXpansion Technology (MXT) or the IBM iSeries compressed disks, that use LZ'77-like coding on small compression units, such as 1-4 KB, and store the compressed data in memory or on disk tracks. Here, when a change of a portion of the compression unit occurs, for example, a L2 cache line, or a 512-byte disk sector, the data is re-compressed and potentially stored in a different location. Knowing the maximum expansion, rather than the average expansion, is an important factor for designing policies for allocation and management of memory or disk space.

1 Introduction

We investigate the maximum increase in number of phrases that results from changing k consecutive symbols in a string that has been parsed using an LZ'77-like algorithm.

We consider a class of compression algorithms that partition a sequence $\mathbf{x} = x_1^n$ into a collection $\mathbf{y}$ of (non-overlapping) variable-length phrases, $y_1, \ldots, y_{N(\mathbf{x})}$, and encode them. Typically, one visualizes the parsing by introducing commas to delimit the phrases. To be more precise, if $y_t = x_a^{a+|y_t|-1}$, then $\mathbf{x}$ contains a substring $x_b^{b+|y_t|-1}$, where $b < a$, satisfying $x_a^{a+|y_t|-1} = x_b^{b+|y_t|-1}$, and $\mathbf{x}$ does not contain a match to a prefix of x_a^n longer than y_t, starting before position a. This parsing

is consistent with the version of LZ'77 [1] described, for example, in [2], where the sliding window is larger than the length of the string $\mathbf{x}$.

We first consider the case $k = 1$. If two phrases $\mathbf{x}$ and $\hat{\mathbf{x}}$ differ in the symbol in position i, in general will have parsings $\mathbf{y}$ and $\hat{\mathbf{y}}$ with a different number of phrases. Our interest is to provide worst case bounds on the number of additional phrases $\|\hat{\mathbf{y}}\| - \|\mathbf{y}\|$ that can result from the singleton change, as a function of the string length and the position of the change. The first main result of the paper, Theorem 2 in Section 3, shows that two strings of length n that differ only in the ith position can yield parsings whose number of phrases differ by as much as $[3(n - i - (3n)^{1/3})]^{2/3}/2$. Because of the symmetry of the problem, the theorem provides bounds to both maximum expansion and maximum contraction resulting from the change. Theorem 2 provides a constructive lower bound and an upper bound. This upper bound is further refined in Theorem 3 of Section 4. The proposed lower bounds require a sufficiently large alphabet; the required minimum size is specified in the theorem, but a rough upper bound is $2^{\lceil [3(n-i+1)]^{1/3} \rceil}$. No specific alphabet size is required for the upper bound. These results are of practical importance: in existing sliding-window implementations of LZ'77, where the alphabet size is 256, the bounds are valid for window sizes of up to 5.7 million symbols, substantially larger than those actually used.

We address the more general situation in which k consecutive symbols are modified in Section 5, where Theorem 4 yields an upper bound and a lower bound. We provide an initial refinement to the analytic result in Section 6, where we also discuss how the upper bound can be further tightened.

In Section 7 we discuss how to extend the results of the paper to the sliding-window LZ'77 algorithm and to the case where the k modified symbols are not necessarily consecutive, and how to approach the question of the maximum expansion of a string of given compressibility.

We are not aware of any earlier work addressing the problem considered here. We point out that Ziv's proof of the convergence of LZ'77 [3], p. 410, employs the idea of approximating a sequence with another sequence that differs from the former in at most a small fraction of the total positions. In contrast, our work is a worst case analysis for finite strings. In addition, Lempel and Ziv [4] have previously introduced the idea of studying packings of all phrases up to a certain length which has some loose resemblance to many of our arguments, like those in Lemma 8 and Lemma 3. Further discussion on Ziv's results can be found in Shields [5], p. 134, including an interesting remark on the author's perception of the potential effect of changes in the Lempel-Ziv parsing of a string. In

a sense, Shields is correct in thinking that changes in a string can dramatically affect the parsing of a string; the possible extent of the effect of the changes is in fact the main topic in this paper. From a basic research point of view, interesting open problems include that of determining features of the probability distribution of the expansion statistic assuming say, a fixed or random number of randomly selected modifications.

From a practical perspective, this problem was motivated by the task of devising management policies for computer systems, such as the IBM Memory eXtension Technology (MXT) [6], that use a shared-dictionary LZ'77-like coding [7], or the IBM iSeries compressed disks, that uses a proprietary version of LZ'77. In these systems, compression is performed on small compression units (1-4 KB) and the compressed data is stored in main memory or on disk tracks.

Ideally, the compression function in a compressed-memory system should be completely invisible to the application running in the computer and, to the extent possible, to the operative system as well. To accomplish this goal, in some designs the hardware assumes almost complete ownership of the memory management problem. The computer system reports to the software that say, twice as much memory is available in the system as physically installed, assuming that the compressibility of the data stored in main memory will be 2:1 or better.

The commitment of available memory made by the hardware to the software is at risk of being broken when the compressibility of the data fails to conform to the original expectations. To help prevent this event from arising, when suitable risk thresholds are exceeded the software alters the compressibility of the data by first storing suitably chosen memory pages in a swapping device (making use of the standard paging mechanisms in virtual memory systems), and then filling with zeros the memory pages thus freed [8].

The above mechanism could in principle ensure a correct system operation if it could be guaranteed that the system will not stop working before the above process is accomplished. The difficulty lies in that compressibility could degrade during the execution of this process. The ensuing problem is termed the Guaranteed Forward Progress problem (Franaszek and Poff, personal communication, see also [9, 10, 11]). A discussion of the general characteristics of the solutions to this problem is outside the scope and target of this paper.

Our work is both motivated by the general scenario described above and our desire to contribute to the basic understanding of properties of compressed streams apparently neglected in the past. Our initial concern with non-probabilistic results is related to the eventual necessity of providing

formal correctness guarantees. Eventually we hope to develop general mathematical tools to answer questions such as

- Can we provide a useful upper bound to the amount of expansion that the memory in a machine can exhibit during in a given period of time?
- What statistics should be tracked to answer the above?
- Can we design compression algorithms with benign expansion properties, yet with little or no compression performance loss?

2 Preliminaries

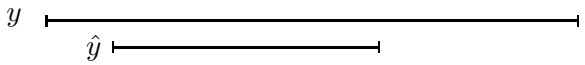
Let $\mathbf{x} \stackrel{\triangle}{=} x_1^n \stackrel{\triangle}{=} [x_1, \ldots, x_n]$ and $\hat{\mathbf{x}} = \hat{x}_1^n$ be two strings, and let $\mathbf{y}$ and $\hat{\mathbf{y}}$ be the corresponding parsings. Let y_r denote the rth phrase of $\mathbf{y}$, and $\hat{y}_s$ denote the sth phrase of $\hat{\mathbf{y}}$. The notation $\|\mathbf{y}\|$ refers to the number of phrases produced by the parsing (or of commas, if we add a comma at the end of each phrase,) while $|y_t|$ is the length of the phrase y_t. If a phrase y corresponds to indexes v through $v + |y| - 1$ in the string $\mathbf{x}$ then we say that $I(y) = [v, \ldots, v + |y| - 1]$. If $\mathcal{Y} = \{y_a, \ldots, y_b\}$, then $I(\mathcal{Y}) \stackrel{\triangle}{=} \bigcup_{y \in \mathcal{Y}} I(y)$. We say that a phrase y overlaps an interval J if $I(y) \bigcap J \neq \emptyset$. We also use the shorthand notation $I(y) \leq a$ (resp. $I(y) \geq a$) if the rightmost (resp. leftmost) index of the interval $I(y)$ is at most a (resp. at least a).

For example, let $\mathbf{x} = ABABCABCAB$, then $\mathbf{y} = A, B, AB, C, ABCAB$, $\|\mathbf{y}\| = 5$, $y_1 = [A]$, $y_5 = [ABCAB]$, $|y_5| = 5$, $I(y_5) = [6, \ldots, 10]$. Also $I(y_5) \leq 10$, $I(y_5) \geq 5$, but $I(y_5)$ is neither ≥ 8 nor ≤ 8.

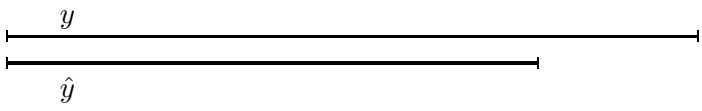
For every phrase y with $I(y) = [v, \ldots, v + |y| - 1]$ let

$$\Phi(y) = \{\hat{y} \in \hat{\mathbf{y}} : I(\hat{y}) = [v + p, \ldots, v + |y| - 1 - q] \quad \text{with} \quad p \geq 0, q > 0\}$$

Note the asymmetric conditions on p and q. Thus $\Phi(y)$ contains all those phrases of the modified parsing $\hat{\mathbf{y}}$ that start and end in $I(y)$ with the added restriction that the rightmost symbol cannot have an index equal to $v + |y| - 1$; graphically, $\Phi(y)$ contains every phrase of the modified parsing of the form



and



As a shorthand for the cardinality of $\Phi(y)$ we will use $\phi(y) = |\Phi(y)|$.

We can easily bound the difference between $\|\hat{\mathbf{y}}\|$ and $\|\mathbf{y}\|$ using the definition of $\phi(\cdot)$:

Theorem 1

$$\|\hat{\mathbf{y}}\| - \|\mathbf{y}\| \leq \sum_{y \in \mathbf{y}} \phi(y). \tag{1}$$

Proof. Every phrase in $\hat{\mathbf{y}} - \bigcup_{y\in\mathbf{y}} \Phi(y)$ contains the rightmost symbol of some phrase of the original parsing $\mathbf{y}$; since phrases are non-overlapping $\left|\hat{\mathbf{y}} - \bigcup_{y\in\mathbf{y}} \Phi(y)\right| \leq \|\mathbf{y}\|$. The theorem follows from the observation that $\|\hat{\mathbf{y}}\| = \sum_{y\in\mathbf{y}} \phi(y) + \left|\hat{\mathbf{y}} - \bigcup_{y\in\mathbf{y}} \Phi(y)\right|$. □

In this paper, we obtain expansion results by exploiting a number of restrictions on the possible values that the function $\phi(\cdot)$ can take on. For example, we will show that if only the ith symbol is changed and $i \in I(y)$, it is always true that $\phi(y) \leq 4$, and that if y is such that $i \notin I(y)$ then $\phi(y) \leq 2$. The following definitions, related to how LZ'77 parses a string, will be used throughout the paper.

Definition 1 *A* previous occurrence (PO) *of a phrase y with $I(y) = [a, \ldots, b]$ (or, more generally, of a substring x_a^b), is a substring x_{a-c}^{b-c} with offset c strictly greater than 0, and satisfying $x_{a-c}^{b-c} = x_a^b$.*

Definition 2 *The* rightmost previous occurrence (RPO) *of a phrase y (or, more generally, of a substring x_a^b) is the previous occurrence having the smallest offset $c > 0$.*

Definition 3 *The interval of positions occupied by the RPO of a phrase y (or, more generally, of a substring x_a^b) is called the* Rightmost Previous Occurrence Interval, or RPOI, *of y. If a string y does not have a RPO, its RPOI is the empty interval.*

Definition 4 *We say that a phrase y* points to *position i if its RPO contains position i. More generally, we say that y points to the interval $\mathcal{I}$ if its RPOI intersects $\mathcal{I}$.*

For example, let $\mathbf{x} = ABABCABDAB$, then $\mathbf{y} = A, B, AB, C, AB, D, AB$. The phrase $y_7 = [AB]$ has $I(y_7) = [9, 10]$, has three POs starting in positions 1, 3, and 6 respectively. Its RPO is the substring x_6^7, and its RPOI is $[6, 7]$.

3 Changing one symbol

We now find bounds to the maximum expansion of a string when only the ith symbol is changed. We use $\mathbf{x}$ and $\mathbf{y}$ to denote the string and its parsing before the ith symbol is changed, and $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ to denote string and parsing after the change.

Theorem 2 *Let* $\mathbf{x}$ *and* $\hat{\mathbf{x}}$ *be two strings of length* n *differing only in the* i*th symbol. Let* $\underline{L} = \max\left\{\mathbb{L} \mid \sum_{\ell=2}^{\mathbb{L}} (\ell \min(\ell, i)) \leq n - i - \mathbb{L} + 1\right\}$*, and let* $\overline{L} = \lceil [3(n-i+1)]^{1/3} \rceil$*. Then*

$$\max_{\mathbf{y}, \hat{\mathbf{y}}[i]} (\|\hat{\mathbf{y}}\| - \|\mathbf{y}\|) \leq \frac{\overline{L}\left(\overline{L}+1\right)}{2} + \lfloor \sqrt{n-i+1} \rfloor + 4 \tag{2}$$

and, if the alphabet size is at least $[\min(i, \underline{L}) + \underline{L}]$*,*

$$\sum_{\ell=2}^{\underline{L}} \min(\ell, i) \leq \max_{\mathbf{y}, \hat{\mathbf{y}}[i]} (\|\hat{\mathbf{y}}\| - \|\mathbf{y}\|). \tag{3}$$

The upper bound of the theorem is universal over all alphabet sizes. The theorem also provides a lower bound that shows that, for large enough alphabet, the upper bound is indeed meaningful, albeit not tight.

We now first provide the reader with an intuitive argument for why the theorem holds, we then introduce the lemma supporting the upper bound, prove the upper bound, and finally prove the lower bound.

Theorem 2 is a consequence of Theorem 1. The first question to be answered is what are the characteristics of the phrases having non-zero $\phi(\cdot)$ in the sum of Inequality (1). Clearly, the (unique) phrase containing position i could have $\phi(\cdot) > 0$, as it is easy to see from simple examples. Also, phrases to the left of position i are not affected by the change.

It turns out that, of the phrases to the right of position i, only those pointing to position i can have $\phi(\cdot) > 0$ (this is established in Lemma 2). A lower bound to $\max_{\mathbf{y}, \hat{\mathbf{y}}[i]} (\|\hat{\mathbf{y}}\| - \|\mathbf{y}\|)$ can then be obtained by bounding from below the maximum number of phrases that could point to position i (as shown in Lemma 8), and by constructing a string with at least this number of phrases that point to position i each of which has $\phi(\cdot) = 1$.

To prove the upper bound, the first step is to bound from above the number of non-zero terms in the sum of Inequality (1) (for instance, by bounding from above the number of phrases pointing to position i, as in Lemma 3). The individual non-zero terms in the sum of Inequality (1) could

be, in principle, arbitrarily large. We show that this is not the case, and that these non-zero terms are at most equal to 2 (Lemma 4).

Not all phrases pointing to position i can have $\phi(\cdot) = 2$. Lemma 5 shows conditions under which $\phi(\cdot) \leq 1$, and its most important consequence is that a phrase y with $\phi(y) = 2$ cannot be preceded by any phrase pointing to position i having length equal to $|y|$ or $(|y| - 1)$. This result yields a bound on the maximum number of phrases not containing position i and with $\phi(\cdot) = 2$ detailed in Lemma 7. The proof of the theorem follows from appropriately combining the mentioned results.

3.1 The Upper Bound of Theorem 2

3.1.1 Supporting Lemma

The following lemma shows that the only phrase overlapping position i can have $\phi(\cdot)$ at most equal to 4.

Lemma 1 *If $i \in I(y)$ then*

$$\phi(y) \leq 4. \tag{4}$$

Proof. Without any loss of generality, we re-index the positions of the symbols in the string $\mathbf{x}$ so that $y = x_0^{|y|-1}$ and that $i \in [0, \ldots, |y| - 1]$. Let $p \geq 1$ be such that $[-p, \ldots, -p + |y| - 1]$ is the RPOI of y. If there is a phrase in $\hat{\mathbf{y}}$ that starts strictly after position $i+p$, then its rightmost border is located at a position greater than or equal to $|y| - 1$ and therefore this phrase does not belong to $\Phi(y)$. Therefore, all we need to do is to argue that at most four phrases of $\hat{\mathbf{y}}$ have starting point in $[0, \ldots, i + p]$.

Since the change occurs at a position greater or equal to zero, the first phrase in $\Phi(y)$ starts at 0 and has a rightmost border greater than or equal to $i - 1$. Now, if a phrase starts somewhere in $[i + 1, \ldots, i + p - 1]$ (this can only happen if $p \geq 2$) then its rightmost border is greater than or equal to $i + p - 1$. Only two other phrases are possible: a phrase that starts at position i and a phrase that starts at position $i + p$. □

From the proof of Lemma 1, it is clear that a necessary condition for $\phi(y) = 4$ is that $i \in \text{RPOI}(y)$, namely, if $\phi(y) = 4$, the position of changed belongs to both the interval and to the RPOI of the phrase y.

The following lemma shows that phrases not overlapping position i and for which $\phi(\cdot)$ is greater than zero must perforce point to position i. The significance of this result, summarized in its

corollary, is that to bound from above the sum in Inequality (1) we need only restrict our attention to phrases pointing to i.

Lemma 2 *If $i \notin I(y)$ and $\phi(y) \geq 1$, then y points to i.*

Proof. Suppose that $i \notin \text{RPOI}(y)$, $i \notin I(y)$ and $\phi(y) \geq 1$. Let c and d be such that $I(y) = [c, \ldots, d]$. By assumption, there is a phrase $\hat{y}$ with $I(\hat{y}) = [c+p, \ldots, d-q]$, where $p \geq 0$ and $q > 0$. Let the RPOI of y be $[c-r, \ldots, d-r]$. Since y does not point to i, we know that $x_{c-r}^{d-r} = \hat{x}_{c-r}^{d-r}$. Also, since $i \notin I(y)$, $x_c^d = \hat{x}_c^d$. Therefore $\hat{x}_{c-r+p}^{d-r}$ is a previous occurrence of $\hat{x}_{c+p}^d$ which contradicts the assumption that the LZ'77 parsing of $\hat{\mathbf{x}}$ yields $\hat{y}$ with $I(\hat{y}) = [c+p, \ldots, d-q]$.

Corollary 1 *Out of the phrases $y \in \mathbf{y}$ with $i \notin I(y)$, only those that point to position i can contribute to an increase in the number of phrases of the parsing of $\mathbf{x}$ when the symbol in position i is changed.*

We now address the question of how many phrases could point to position i. The following lemma provides us with an upper bound.

Lemma 3 ***(Packing Lemma A)*** *Let $\overline{L} = \lceil [3(n-i+1)]^{1/3} \rceil$. The maximum number N_1 of phrases of $y \in \mathbf{y}$ such that $i \notin I(y)$ and y points to position i is bounded from above by*

$$N_1 < \frac{\overline{L}\left(\overline{L}+1\right)}{2}. \tag{5}$$

Proof. Construct a pseudo parsing as follows. Add one comma between position i and position $i+1$, skip 2 symbols, and add a comma, repeat once, skip 3 symbols and add a comma, repeat twice, etc. The resulting pseudo-parsing of $\mathbf{x}$ has, to the right of position i, 1 pseudo-phrase of length 1, 2 pseudo-phrases of length 2, 3 pseudo-phrases of length 3, etc. Note that there are at most ℓ phrases of length ℓ that can point to i. Therefore, for no string of length n the LZ'77 parsing can produce more phrases that point to position i than the pseudo-parsing algorithm does, because the pseudo-parsing fills the available space with all available short phrases.

The pseudo-phrases of length $\leq \ell$ occupy $\sum_{j=1}^{\ell} j^2$ positions. Hence, the longest pseudo-phrase contains at most $k^* = \min\{k \mid \sum_{j=1}^{k} j^2 \geq (n-i)\}$ symbols. To compute k^*, we note that $\sum_{j=1}^{k} j^2 = k(k+1)(2k+1)/6$. We then solve the equation $x(x+1)(2x+1)/6 = n-i$, or

$$f(x) = x^3 + \frac{3}{2}x^2 + \frac{1}{2}x - 3(n-i) = 0, \tag{6}$$

which has a unique positive root for $n-i>0$, because $f^{'}(x)>0$ for every $x>0$, and $f(0)<0$[1]. For every $x>0$, $f(x)>g(x)=x^3+\sqrt{3/2}x^2+x/2-3(n-i)$, because $f(x)-g(x)=\left(3/2-\sqrt{3/2}\right)x^2>0,\ \forall x>0$. Hence, the positive solution of $f(x)=0$ is strictly smaller than the positive solution of $g(x)=0$. Because the square of the coefficient of the x^2 term equals 3 times the coefficient of the x^1 term, the equation $g(x)=0$ has an easily computed unique real solution, $x^*=-\frac{1}{3}\sqrt{\frac{3}{2}}+\left[\left(\frac{3}{2}\right)^{3/2}\frac{1}{27}+3(n-i)\right]^{1/3}=-\sqrt{\frac{1}{6}}+\left[\frac{1}{6^{3/2}}+3(n-i)\right]^{1/3}$, which is less than $[3+3(n-i)]^{1/3}$ and can be bounded from above as $\lceil[3(n-i+1)]^{1/3}\rceil\stackrel{\triangle}{=}\overline{L}$. The maximum number of phrases that have length $\overline{L}$ or less is then $\sum_{j=1}^{\overline{L}}j=\overline{L}(\overline{L}+1)/2$. □

Note: This proof can be shortened by noting that $f(\overline{L})>0$ and therefore $k^*\leq\overline{L}$. We have chosen to present the longer proof in order to motivate our choice for the value of $\overline{L}$.

The following lemma shows that the contribution of individual phrases pointing to and not containing position i to the sum in Inequality (1) is at most 2.

Lemma 4 *If y points to i and $i\notin I(y)$, then $\phi(y)\leq 2$.*

Proof. The statement is trivial if $|y|\leq 3$. Let m, a, and b be such that $I(y)=[m-a,\ldots,m+b]$ and $\mathrm{RPOI}(y)=[i-a,\ldots,i+b]$, with $a,b\geq 0$. Let $\hat{y}_s,\ldots,\hat{y}_{s+t}$ be all the phrases of $\hat{\mathbf{y}}$ starting in the interval $[m-a,\ldots,m+b-1]$. If one of these phrases, call it $\hat{y}_r$, starts after position m, say $I(\hat{y}_r)=[m+c,\ldots,m+d]$, $d\geq c>0$, then $d\geq b$, because $\hat{x}_{m+c}^{m+b}=x_{m+c}^{m+b}$, $\hat{x}_{i+c}^{i+b}=x_{i+c}^{i+b}$, and $\hat{x}_{i+c}^{i+b}$ is a previous occurrence of $\hat{x}_{m+c}^{m+b}$. Hence, $r=s+t$ and $\hat{y}_r\notin\Phi(y)$.

We now show that there are at most 2 phrases, $\hat{y}_s$ and $\hat{y}_{s+1}$, that start in the closed interval $[m-a,\ldots,m]$. The phrase $\hat{y}_s$ (if it exists) must have $I(\hat{y}_s)=[m-a+e,\ldots,m-1+f]$ with $e\geq 0$ (by definition) and $f\geq 0$ because $\hat{x}_{m-a}^{m-1}=x_{m-a}^{m-1}$, $x_{i-a}^{i-1}=\hat{x}_{i-a}^{i-1}$, and $\hat{x}_{i-a}^{i-1}$ is a previous occurrence of $\hat{x}_{m-a}^{m-1}$. Therefore, $\hat{y}_{s+1}$ (if it exists) must start after or at position m. If the former is true, then $\hat{y}_{s+1}=\hat{y}_r\notin\Phi(y)$ and therefore $\phi(y)=1$. If the latter is true, then $\phi(y)\leq 2$ since $\hat{y}_{s+1}$ may or may not belong to $\Phi(y)$ and $\hat{y}_{s+2}\notin\Phi(y)$. □

The following lemma shows conditions under which $\phi(y)\leq 1$. We use this Lemma in the proof of both Theorems 2 and 4 and therefore we state it for the case where $k\geq 1$ consecutive symbols are changed. Let $\mathcal{I}$ denote the interval $[i,\ldots,i+k-1]$ where the changes occur.

[1] A reviewer pointed out that the proof could be concluded here by noting that $f(\overline{L})>0$. We provide a longer proof to show how $\overline{L}$ arises, since we refer to it in the appendix.

Lemma 5 *Let y_t be a phrase of* $\mathbf{y}$ *having length $a+b+k$ with $I(y_t) = [m-a, \ldots, m+k-1+b]$, $\mathcal{I} \bigcap I(y_t) = \emptyset$, and RPOI$(y_t) = [i-a, \ldots, i+k-1+b]$. Then $\phi(y_t) \leq 1$ if any of the following conditions hold:*

Condition 1. *There exists $p, q \geq 0$ and $i < j < m$ such that $\mathcal{I} \bigcap [j-a-p, \ldots, j+k-1+q] = \emptyset$ and $x_{j-a-p}^{j+k-1+q} = x_{i-a-p}^{i+k-1+q}$.*

Condition 2. *There exists $p, q \geq 0$ and $i < j < m$ such that $\mathcal{I} \bigcap [j-p, \ldots, j+k-1+b+q] = \emptyset$ and $x_{j-p}^{j+k-1+b+q} = x_{i-p}^{i+k-1+b+q}$.*

Hence, if the RPOIs of two phrases contain the changed interval and intersect in a way that neither is contained in the other with padding at both ends, the rightmost of the two phrases cannot have $\phi(\cdot) \geq 2$.

Proof.

We first establish the following facts:

Fact 1. If Condition 1 holds, b must be greater than 0, $\hat{x}_{m-a}^{m+k-1}$ has at least one previous occurrence in $\hat{\mathbf{x}}$, and $\hat{x}_{m+k}^{m+k-1+b}$ has at least one previous occurrence in $\hat{\mathbf{x}}$.

Fact 2. If Condition 2 holds, a must be greater than 0 $\hat{x}_{m-a}^{m-1}$ has at least one previous occurrence in $\hat{\mathbf{x}}$, and $\hat{x}_{m}^{m+k-1+b}$ has at least one previous occurrence in $\hat{\mathbf{x}}$.

First, restrict the attention to Fact 1. We first note that b must clearly be greater than zero otherwise x_{m-a}^{m+k-1} would be copied from x_{j-a}^{j+k-1}; since $x_{i-a}^{i+k-1} = x_{j-a}^{j+k-1} = x_{m-a}^{m+k-1}$, $\mathbf{x}$ and $\hat{\mathbf{x}}$ differ only in the $\mathcal{I}$ positions, and $\mathcal{I} \bigcap I(y_t) = \emptyset$, we conclude that $x_{j-a}^{j+k-1} = \hat{x}_{j-a}^{j+k-1}$, $x_{m-a}^{m+k-1} = \hat{x}_{m-a}^{m+k-1}$. Therefore $\hat{x}_{j-a}^{j+k-1}$, is a PO (see Definition 1) of $\hat{x}_{m-a}^{m+k-1}$. Also, $x_{i+k}^{i+k-1+b} = \hat{x}_{i+k}^{i+k-1+b}$, $x_{i+k}^{i+k-1+b} = x_{m+k}^{m+k-1+b}$, and $\hat{x}_{m+k}^{m+k-1+b} = x_{m+k}^{m+k-1+b}$, which shows that $\hat{x}_{i+k}^{i+k-1+b}$ is a PO of $\hat{x}_{m+k}^{m+k-1+b}$.

Then, consider Fact 2. Clearly a must be > 0, otherwise $x_{m}^{m+k-1+b}$ would be copied from $x_{j}^{j+k-1+b}$; The same arguments used in establishing Fact 1 show that $\hat{x}_{j}^{j+k-1+b}$ is a PO of $\hat{x}_{m}^{m+k-1+b}$; Also $\hat{x}_{i-a}^{i-1}$ is a PO of $\hat{x}_{m-a}^{m-1}$.

Suppose now that Condition 1 holds, and let $\hat{y}_u \in \Phi(y_t)$. If $\hat{y}_u$ has starting point at or after position $m+k$, then by Fact 1 its rightmost border is at least $m+k-1+b$ and by definition, $\hat{y}_u \notin \Phi(y_t)$, a contradiction. Thus the index of the starting point of $\hat{y}_u$ is at most $m+k-1$, and by Fact 1, the index of the rightmost border is at least $m+k-1$. Since as discussed, any phrase that starts after $m+k-1$ cannot be in $\Phi(y_t)$, $\phi(y_t) \leq 1$ and the Lemma is proved for this case.

We can use an analogous argument to show that $\phi(y_t) \leq 1$ when Condition 2 is true. □

Lemma 5 has a fundamental consequence:

Lemma 6 *Let y_s be a phrase of length $|y_s|$, such that $i \in RPOI(y)$ and $i \notin I(y_s)$. Then, for every phrase y_t with $t > s$ of length $|y_t| \leq |y_s| + 1$, $\phi(y_t) \leq 1$.*

Proof. Let the RPOI of y_s be $[i-a, \ldots, i+b]$ with $a, b \geq 0$. Let RPOI of y_t be $[i-c, \ldots, i+d]$, with $c, d \geq 0$. Note that $a \geq c$ or $b \geq d$, because $|y_s| + 1 = (a+b+1) + 1 \geq (c+d+1) = |y_t|$. If $a \geq c$, Condition 1 of Lemma 5 holds. If $b \geq d$, Condition 2 of Lemma 5 holds. □

Lemma 6 implies that not all phrases counted in Lemma 3 can have $\phi(\cdot) = 2$. A bound on the number of phrases not containing position i and with $\phi(\cdot) = 2$ is provided by the following lemma.

Lemma 7 *The number N_2 of phrases $y \in \mathbf{y}$ with $\phi(y) = 2$ and $i \notin I(y)$ is bounded from above by*

$$N_2 \leq \lfloor \sqrt{n-i+1} \rfloor \tag{7}$$

Proof. We use a packing argument. Note that a phrase y with $\phi(y) = 2$ and $i \notin I(y)$ has length 3 or more. Lemma 6 shows that there can exist only one phrase of length 3 that points to position i, does not contain position i, and has $\phi(\cdot) = 2$. If this phrase exists, then we can have at most one phrase of length 5 that has $\phi(\cdot) = 2$. The same lemma also prevents any phrase of length 4 from having $\phi(\cdot) > 1$. The cumulative length of all phrases that point to position i, do not contain position i, and have $\phi(\cdot) = 2$ cannot be more than $n - i$. The maximum number N_2 of such phrases is therefore bounded from above by $\min\left\{k | \sum_{j=1}^{k}(2j+1) \geq (n-i)\right\}$. Recall that $\sum_{j=1}^{k}(2j+1) = k^2 + 2k$. The solutions of the equation $x^2 + 2x - (n-i) = 0$ are $-1 \pm \sqrt{n-i+1}$. Therefore $N_2 \leq \lfloor \sqrt{n-i+1} \rfloor$. Any other packing would substitute a longer phrase for a shorter phrase and cannot yield more phrases with $\phi(\cdot) > 1$. □

Proof of the Upper Bound

Our starting point is Theorem 1. There is exactly one phrase $y \in \mathbf{y}$ for which $i \in I(y)$, and this phrases has $\phi(y) \leq 4$ by Lemma 1, Inequality (4).

We now consider phrases y such that $i \notin I(y)$. Corollary 1 ensures that, out of these phrases, the only ones that can contribute to the expansion (in other words, contribute a positive value of $\phi(y)$ in the summation in Theorem 1), must point to i. Lemma 3 bounds from above the maximum number of phrases pointing to i by Inequality (5), namely, by $\left[\overline{L}\left(\overline{L}+1\right)\right]/2$, where $\overline{L} = \lceil [3(n-i+1)]^{1/3} \rceil$.

By Lemma 4, if $i \notin I(y)$ then $\phi(y) \leq 2$; the maximum number of phrases for which $\phi(\cdot) = 2$ is bounded from above in Lemma 7 by $\lfloor \sqrt{n-i+1} \rfloor$ (Inequality 7).

The upper bound is then obtained by combining Inequalities (5), (4), and (7). □

3.2 The Lower Bound of Theorem 2

The proof of the lower bound is constructive, and intimately relies on an additional packing lemma, which we introduce first. This lemma bounds from below the maximum number of phrases that can point to position i in the parsing $\mathbf{y}$ of a string $\mathbf{x}$ having length n.

Lemma 8 ***(Packing Lemma B)*** *Let* $\underline{L} = \max\left\{\mathbb{L} \mid \left(\sum_{\ell=2}^{\mathbb{L}} \ell \min(\ell, i)\right) \leq n - i - \mathbb{L} + 1\right\}$ *If the alphabet size is at least* $\min(i, \underline{L}) + \underline{L}$*, the maximum number* N_0 *of phrases of* $\mathbf{y}$ *pointing to position* i *is bounded from below by*

$$N_0 \geq \sum_{\ell=2}^{\underline{L}} \min(\ell, i). \tag{8}$$

If $i \geq (3(n+2))^{\frac{1}{3}}$*, then* $\underline{L} = \lfloor 3[n - i - (3(n+2))^{\frac{1}{3}}]^{1/3} \rfloor$.

Proof. We construct a string $\mathbf{x}$ that satisfies Inequality 8, using the following packing algorithm. Let the substring $x_{i-\min(i,\underline{L})+1}^{i+\underline{L}-1}$ consist of distinct symbols of the alphabet. There is at least one unused alphabet symbol, which we denote by $\tilde{\alpha}$, that we use to fill the positions to the left of $\max\{i - \underline{L} + 1, 1\}$. We now consider only substrings containing position i. There are $\min(i, 2)$ such substrings having length 2, $\min(i, 3)$ having length 3, etc. Starting from position $i + \underline{L}$ copy first the substrings of length 2 from right to left into positions $i + \underline{L}, i + \underline{L} + 1, \ldots$, then all the substrings of length 3 from left to right, and so on, until all the substrings of length $\underline{L}$ containing position i have been copied (note the alternating order of the copy). The definition of $\underline{L}$ ensures that there is sufficient space in $\mathbf{x}$ to contain all these copies. If there is additional space at the beginning or at the end of $\mathbf{x}$, fill it with copies of $\tilde{\alpha}$. Parse $\mathbf{x}$ into $\mathbf{y}$. The first $i-1$ symbols are irrelevant, and the symbols in the interval $[i, \ldots, i + \underline{L}]$ parse into phrases of length 1. Note that the last symbol of each copied substring and the first symbol of the next copied substring appear only once as a pair in the entire $\mathbf{x}$. Hence, when parsing $\mathbf{x}$ these symbols are always separated by a comma. Also, each copied substring containing position i by construction has a unique previous occurrence, namely, the substring from which it was copied. Hence, $\mathbf{y}$ contains exactly $\sum_{\ell=2}^{\underline{L}} \min(\ell, i)$ phrases pointing to position i. □

Note: if $i > i^*$, where i^* is 2 plus the ceiling of the positive solution of the equation $x^3 - 3(n-x+1) = 0$, then the technique used in the next lemma can simplify this lower bound to to $K(K+1)/2$, where $K = [3(n-i-i^*+1)]^{1/3} + 2$.

Proof of the Lower Bound

Lemma 8 constructively shows that the maximum number N_1 of phrases pointing to position i is bounded from below by $\sum_{\ell=2}^{\underline{L}} \min(\ell, i)$, if the alphabet size is sufficiently large, say $\geq [\min(i, \underline{L}) + \underline{L}]$ (if the alphabet size is 256, and $i > 127$, n can be as large as 5.7 million, a substantial number compared to the window sizes used in practical LZ'77 implementations). We use the string $\mathbf{x}$ and the "unused symbol" $\tilde{\alpha}$ defined in Lemma 8. We build $\hat{\mathbf{x}}$ by changing the ith symbol of $\mathbf{x}$ to $\tilde{\alpha}$. The string $\mathbf{x}$ is carefully crafted to ensure that each pair of symbols straddling commas in the parsing (i.e., the last symbol of a phrase and the first symbol of the following phrase) appear only once in both $\mathbf{x}$ and $\hat{\mathbf{x}}$. Hence, each comma of $\mathbf{y}$ has a corresponding comma in $\hat{\mathbf{y}}$, and for each phrase y in $\mathbf{y}$ there is a phrase $\hat{y}(y) \in \Phi(y)$ in $\hat{\mathbf{y}}$ starting at $\min I(y)$. Additionally, $|y| > 1$ implies $|\hat{y}(y)| < |y|$, because y has a unique PO and $i \in \text{RPOI}(y)$. The uniqueness of the PO follows from the following considerations: if $i - \underline{L} + 1 > 0$, the symbol used to fill the positions to the left of $i - \underline{L} + 1$ does not appear in phrases pointing to position i. The phrase y has a unique PO intersecting the interval $[i - \underline{L} + 1, \ldots, i + \underline{L} - 1]$ because the symbols in this interval are different. No phrase to the right of $[i - \underline{L} + 1, \ldots, i + \underline{L} - 1]$ and to the left of y is equal to y by construction. It remains to be shown that y cannot be copied from any substring of $\mathbf{x}$ starting to the to the right of $[i - \underline{L} + 1, \ldots, i + \underline{L} - 1]$ and to the left of y, and perforce intersecting at least two adjacent phrases having a PO that contains i (where the rightmost of such phrases could be y itself). But this is not possible, because for each pair of adjacent phrases each having a PO that contains i, the last symbol of the first phrase and the first symbol of the second phrase appear as a pair only once in the entire string $\mathbf{x}$, as discussed in the proof of Lemma 8. Hence, each phrase $y \in \mathbf{y}$ with $i \in \text{RPOI}(y)$ has $\phi(y) \geq 1$. Lemma 6 ensures that $\phi(y) \leq 1$ because each phrase $y \in \mathbf{y}$ with $i \in \text{RPOI}(y)$ and $|y| > 2$ is preceded by a phrase $y'(y) \in \mathbf{y}$ satisfying $\left|y'(y)\right| = |y| - 1$ and $i \in \text{RPOI}(y')$. We conclude that each phrase $y \in \mathbf{y}$ with $|y| \geq 2$ has $\phi(y) = 1$.

4 Result Refinement

The upper bound of Theorem 2 was obtained by bounding the maximum number of phrases with $\phi(\cdot) = 1$, bounding the maximum number of phrases with $\phi(\cdot) = 2$, and adding the bounds to the maximum contribution of the phrase overlapping position i. Clearly, the bound can be refined, as the following theorem shows.

Theorem 3 *Under the conditions of Theorem 2,*

$$\max_{\mathbf{y},\hat{\mathbf{y}}[i]} (\|\hat{\mathbf{y}}\| - \|\mathbf{y}\|) \leq \frac{M\,(M+1)}{2} + 4, \tag{9}$$

where $M = \left\lceil \left[3\left((n + \left\lfloor\sqrt{n-i+1}\right\rfloor) - i + 1\right)\right]^{1/3} \right\rceil$.

We know from Lemma 1 how to bound the contribution of a phrase y with $i \in I(y)$ to the sum in Inequality (1). For the refined upper bound we construct, like in the proof of Lemma 3, a fictitious string, this time of an appropriate length $m \geq n$, and a pseudo-parsing $\mathbf{z}$ containing all pseudo-phrases of length 2 up to M that have distinct RPOIs containing position i. We show that this pseudo-parsing contains a number of phrases that is larger than $\sum_{y:I(y)>i} \phi(y)$ for any pair of strings $\mathbf{x}$ and $\hat{\mathbf{x}}$ of length n and differing only in the symbol in position i. We consider any string $\mathbf{x}$, its parsing $\mathbf{y}$, change the ith symbol, and look at the parsing $\hat{\mathbf{y}}$ of the resulting string $\hat{\mathbf{x}}$. We divide the phrases pointing to and not overlapping position i into $Y_1 \stackrel{\triangle}{=} \{y \mid \phi(y) \leq 1\}$ and $Y_2 \stackrel{\triangle}{=} \{y \mid \phi(y) = 2\}$. We further divide these sets into Y_1^S and Y_2^S, containing "short" phrases with length $\leq M$, and Y_1^L and Y_2^L containing "long" phrases with length $> M$. If m is sufficiently large, then we can construct a correspondence between phrases of Y_1^S and phrases in the pseudo parsing $\mathbf{z}$ that point to position i (call this set Z) and between phrases of Y_2^S and *pairs* of phrases in Z that are not already paired with phrases in Y_1^S. Finally, we show that the number of phrases in Z that have not yet been paired is larger than or equal to the size of Y_1^L plus twice the size of Y_2^L.

We briefly remark that Theorem 3 does not produce a bound that is uniformly better than that of Theorem 2. More specifically, for $n - i = 1,\ 7,\ 8$, and $n - i$ from 17 to 20, the approximations introduced by the ceilings and floors in the equations cause Theorem 3 to produce a worse upper bound than Theorem 2 and for $n - i = 35$ to 40 the two bounds are identical. For large values of $n - i$, Theorem 3 yields a substantially tighter bound than Theorem 2.

We now state and prove two supporting lemmas, and subsequently provide a formal proof of the theorem. First, we note that there is a trade off between the existence of phrases in Y_2^S and in

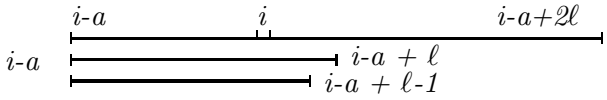
Y_1^S. In particular, given a phrase y_t in Y_2^S, we can look at its RPOI and determine that there are two different intervals containing the changed position i, call them $I_1(y_t)$ and $I_2(y_t)$, that cannot be RPOIs of phrases, in particular of phrases of Y_1^S.

Lemmas 9 and 10 identify these intervals for y_t having odd and even length respectively, and their structures show that $I_1(y_t)$ and $I_2(y_t)$ are different for different y_t's, as discussed later.

The sum of the lengths of $I_1(y_t)$ and $I_2(y_t)$ is at most $|y_t|+1$. This implies that we can compare the original parsing $\mathbf{y}$ to that of a pseudo-string having length at most $|\mathbf{x}|+1$, whose parsing $\mathbf{z}$ contains exactly the same phrases as $\mathbf{y}$ except y_t, and in addition contains two pseudo-phrases $\in Y_1^S$ pointing to $I_1(y_t)$ and $I_2(y_t)$. Then, $\sum_y \phi(y)$ is the same for both pseudo-phrases.

Lemma 9 *Let $\phi(y_t) > 1$, $|y_t| = 2\ell+1$, $\ell \geq 1$, $I(y_t) = [m-a, \ldots, m-a+2\ell]$, $0 < a < 2\ell$, and $RPOI(y_t) = [i-a, \ldots, i-a+2\ell]$.*

1. *If $a < \ell$, let $I_1(y_t) \stackrel{\triangle}{=} [i-a, \ldots, i-a+\ell]$ and $I_2(y_t) \stackrel{\triangle}{=} [i-a, \ldots, i-a+\ell-1]$. Then for every $j < m-a$ such that $i \notin [j, \ldots, j+\ell]$, $x_j^{j+\ell} \neq x_{i-a}^{i-a+\ell}$, and for every $j < m-a$ such that $i \notin [j, \ldots, j+\ell-1]$, $x_j^{j+\ell-1} \neq x_{i-a}^{i-a+\ell-1}$.*



2. *If $a > \ell$, let $I_1(y_t) \stackrel{\triangle}{=} [i-a+\ell, \ldots, i-a+2\ell]$ and $I_2(y_t) \stackrel{\triangle}{=} [i-a+\ell+1, \ldots, i-a+2\ell]$. Then for every $j < m-a$ such that $i \notin [j, \ldots, j+\ell]$, $x_j^{j+\ell} \neq x_{i-a+\ell}^{i-a+2\ell}$ and for every $j < m-a-1$ such that $i \notin [j+1, \ldots, j+\ell]$, $x_{j+1}^{j+\ell} \neq x_{i-a+\ell+1}^{i-a+2\ell}$.*

3. *If $a = \ell$, let $I_1(y_t) \stackrel{\triangle}{=} [i-a, \ldots, i-a+\ell]$, $I_2(y_t) \stackrel{\triangle}{=} [i-a+\ell, \ldots, i-a+2\ell]$. Then, for every $j < m-a$ such that $i \notin [j, \ldots, j+\ell]$, $x_j^{j+\ell} \neq x_{i-a}^{i-a+\ell}$ and for every $j < m-a+\ell$ such that $i \notin [j, \ldots, j+\ell]$, $x_j^{j+\ell} \neq x_{i-a+\ell}^{i-a+2\ell}$.*

Proof.

Let $\phi(y_t) = b+1$ with $b \geq 1$. We prove the three parts of the lemma by contradiction.

1. Let $a < \ell$ and assume that there exists j satisfying the conditions of part 1, namely, $x_j^{j+\ell} = x_{i-a}^{i-a+\ell}$ or $x_j^{j+\ell-1} = x_{i-a}^{i-a+\ell-1}$. Thus $x_j^{j+\ell-1} = x_{i-a}^{i-a+\ell-1} = x_{i-m}^{i-a+m-1}$. Lemma 5 states that $x_{j-a}^{j-a+\ell-1} = x_{i-m}^{i-a+m-1}$ ensures that $\phi(y_t) = 1$, which contradicts the assumption.

2. Let $a > \ell$ and assume that there exists j satisfying the conditions of part 2, namely $x_j^{j+\ell} = x_{i-a+\ell}^{i-a+2\ell}$ or $x_{j+1}^{j+\ell} = x_{i-a+\ell+1}^{i-a+2\ell}$. Thus, $x_{j+1}^{j+\ell} = x_{i-a+\ell+1}^{i-a+2\ell} = x_{m-a+\ell+1}^{i-a+2\ell}$. Again, Lemma 5 states that $x_{j-a+\ell+1}^{j-a+2\ell} = x_{m-a+\ell+1}^{i-a+2\ell}$ ensures that $\phi(y_t) = 1$, which contradicts the assumption.

3. The third part of the lemma is an immediate consequence of Lemma 5.

Lemma 10 is the equivalent of Lemma 9 for the case in which $|y_t|$ is even.

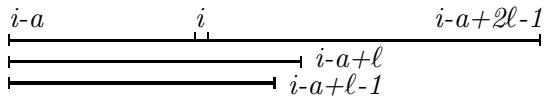
Lemma 10 *Let $\phi(y_t) > 1$, $|y_t| = 2\ell$, $\ell \geq 1$, $I(y_t) = [m-a, \ldots, m-a+2\ell-1]$ and RPOI$(y_t) = [i-a, \ldots, i-a+2\ell-1]$. Consider any j satisfying the following conditions:*

Condition 1*: $j < m-a$,*

Condition 2*: $i \notin [j, \ldots, j+\ell]$.*

The following statements hold:

1. *If $a \leq \ell$, let $I_1(y_t) = [i-a, \ldots, i-a+\ell]$, and $I_2(y_t) = [i-a, \ldots, i-a+\ell-1]$. Then $x_j^{j+\ell} \neq x_{i-a}^{i-a+\ell}$, and, trivially, $x_j^{j+\ell-1} \neq x_{i-a}^{i-a+\ell-1}$.*



2. *If $a > \ell$, let $I_1(y_t) = [i-a+\ell, \ldots, i-a+2\ell]$, and $I_2(y_t) = [i-a+\ell+1, \ldots, i-a+2\ell]$. Then $x_j^{j+\ell} \neq x_{i-a+\ell}^{i-a+2\ell}$ and, trivially, $x_{j+1}^{j+\ell} \neq x_{i-a+\ell+1}^{i-a+2\ell}$,*

Proof. The proof is analogous to that of Lemma 9 and is left to the interested reader.

We remark that different phrases not containing position i and with $\phi(\cdot) = 2$ are associated with different intervals $I_1(\cdot)$ and $I_2(\cdot)$. In particular, let y_s and y_t be two such phrases, with $s < t$. Then RPOI(y_t) contains RPOI(y_s) with padding at both ends (this is an immediate consequence of Lemma 5). But the intervals $I_1(y_s)$ and $I_2(y_s)$ are subintervals of RPOI(y_s) aligned with one of its end points, while the intervals $I_1(y_t)$ and $I_2(y_t)$ are subintervals of RPOI(y_t) aligned with one of its end points. Hence these intervals are distinct.

4.1 Proof Theorem 3

Let $m = n + \left\lfloor \sqrt{n-i+1} \right\rfloor$, where $\left\lfloor \sqrt{n-i+1} \right\rfloor$ is the maximum number of phrases with $\phi(\cdot) = 2$ that do not overlap position i in a string $\mathbf{x}$ of length n, and let $M \stackrel{\triangle}{=} \lceil [3(m-i+1)]^{1/3} \rceil$. Here M is analogous to L in Theorem 2, but is computed for a string of length m rather than n.

As in the proof of Lemma 3, construct a fictitious string, this time of length m, and a pseudo-parsing $\mathbf{z}$ containing all pseudo-phrases of length 2 up to M that have distinct RPOIs containing position i. Call Z the set of such pseudo-phrases. We show that this pseudo-parsing contains a number of phrases pointing to position i that is larger than $\sum_{y:I(y)>i}\phi(y)$ for any pair of strings $\mathbf{x}$ and $\hat{\mathbf{x}}$ of length n and differing only in the symbol in position i. The structure of $\mathbf{z}$ and Lemma 5 ensure that the pseudo-phrases in $\mathbf{z}$ have at most $\phi(\cdot)=1$, and therefore we will be able to use $|Z|$ in the new upper bound.

To prove the statement, consider any string $\mathbf{x}$, its parsing $\mathbf{y}$, and change the ith symbol to obtain $\hat{\mathbf{x}}$ parsed as $\hat{\mathbf{y}}$. Consider all phrases of $\mathbf{y}$ pointing to position i and not containing position i. Divide these phrases into $Y_1=\{y \mid \phi(y)\leq 1\}$ and $Y_2=\{y \mid \phi(y)=2\}$. Divide Y_1 into Y_1^S, containing "short" phrases having length $\leq M$ and Y_1^L, containing "long" phrases having length $>M$. Similarly, divide Y_2 into Y_2^S, containing phrases having length $\leq 2M-1$, and Y_2^L, containing phrases having length $\geq 2M$.

We prove the theorem by showing that

$$\|Y_1^S\|+2\|Y_2^S\|+\|Y_1^L\|+2\|Y_2^L\|\leq\|Z\|. \tag{10}$$

To prove Equation 10 we construct a correspondence between phrases in Y_1^S and phrases in Z, and between phrases in Y_2^S and pairs of phrases in Z not already paired with phrases in Y_1^S. Then we show that the number of remaining phrases in Z is greater than or equal to $\|Y_1^L\|+2\|Y_2^L\|$.

By construction, the phrases of Z have different (pseudo) RPOI, and the same holds for the phrases of $\mathbf{y}$. We then say that a phrase of Z corresponds to a phrase of Y_1^S if they have the same RPOI. This is a 1:1 correspondence. For each phrase of Y_2^S, Lemmas 9 and 10 ensure that two different intervals containing i exist that are not RPOIs of phrases of Y_1^S. Corollary 6 and Lemma 5 ensure that these intervals are different for different phrases of Y_2^S. We have therefore partitioned the set of pseudo-phrases Z into Z_P, the pseudo-phrases in correspondence with phrases in Y_1^S or Y_2^S, and $Z\setminus Z_P$. From Lemmas 9 and 10 it follows that the length of a phrase of Y_2^S is not longer than the sum of the lengths of the corresponding pseudo-phrases of Z minus 1, namely

$$\sum_{z\in Z_P}|z|\leq\sum_{y\in Y_1^S}|y|+\sum_{y\in Y_2^S}(|y|+1). \tag{11}$$

We now deal with Y_1^L and Y_2^L. Note the following inequalities

$$\sum_{z\in Z}|z| \overset{(a)}{\geq} m-i+1$$

$$\begin{aligned} &\overset{(b)}{=} n-i+1+\lfloor\sqrt{n-i+1}\rfloor \\ &\overset{(c)}{\geq} \sum_{y\in Y_1^S}|y|+\sum_{y\in Y_2^S}(|y|+1)+\sum_{y\in Y_1^L}|y|+\sum_{y\in Y_2^L}(|y|+1), \end{aligned} \tag{12}$$

where Inequality (a) follows from the construction of the reference phrase, Equality (b) follows from the definition of m, and Inequality (c) follows from the fact that the number of phrases in $\mathbf{y}$ with $\phi(\cdot)=2$ is bounded from above by $\lfloor\sqrt{n-i+1}\rfloor$ and from the fact that the sum of the lengths of the phrases of $\mathbf{y}$ pointing to position i but not containing i is at most $n-i+1$.

From Inequalities (11) and (12) it follows immediately that

$$\sum_{z\in Z\setminus Z_P}|z| \geq \sum_{y\in Y_1^L}|y|+\sum_{y\in Y_2^L}(|y|+1). \tag{13}$$

Note the following set of inequalities:

$$\begin{aligned} \|Z\setminus Z_P\|\cdot M &= \sum_{z\in Z\setminus Z_P} M \\ &\overset{(a)}{\geq} \sum_{z\in Z\setminus Z_P}|z| \\ &\overset{(b)}{\geq} \sum_{y\in Y_1^L}|y|+\sum_{y\in Y_2^L}(|y|+1) \\ &\overset{(c)}{>} \left\|Y_1^L\right\| M+\left\|Y_2^L\right\|\cdot 2\cdot M, \end{aligned} \tag{14}$$

where Inequality (a) follows from the fact that $\mathbf{z}$ contains phrases of length at most M, Inequality (b) is Inequality (13), and Inequality (c) follows from the definition of Y_1^L (which contains phrases of length $> M$) and Y_2^L (which contains phrases of length $> 2M-1$). From Inequality (14) we conclude immediately that $\|Z\setminus Z_P\| \geq \left\|Y_1^L\right\| + 2\cdot\left\|Y_2^L\right\|$. □

The only other possible contribution to expansion can come from a phrase y with $i\in I(y)$, and is bounded by Lemma 1. The sum of the two bounds gives the bound in the theorem.

5 Changing k consecutive symbols

We now capitalize on the results of the previous sections to compute the maximum expansion of a string of length n when k consecutive symbols starting from position i are changed. As we did for the previous results, we consider separately the phrases overlapping the k changed symbols and the remaining phrases. Recall that $\mathcal{I}=[i,\ldots,i+k-1]$ denotes the interval where $\mathbf{x}$ and $\hat{\mathbf{x}}$ differ.

We first note that Corollary 1 easily extends to k consecutive changes (the proof is left to the interested reader).

Corollary 2 *If $y \in \mathbf{y}$ is such that $\mathcal{I} \bigcap I(y) = \emptyset$ and $\phi(y) \geq 1$, then y points to $\mathcal{I}$.*

The following definitions are needed in the main result of this section.

Definition 5 *A subset $\{y_{j_1}, \ldots, y_{j_m}\}$ of the parsing $\mathbf{y}$* covers *$\mathcal{I}$ if no phrase y_{j_h} intersects $\mathcal{I}$, the RPOI of each phrase y_{j_h} intersects the interval, and the union of the RPOIs of the phrases contains $\mathcal{I}$. The set of RPOIs of the phrases y_{j_h} is called a* covering *of the interval $\mathcal{I}$.*

A covering of an interval need not exist. However, we can always construct a *generalized covering.*

Definition 6 *A set of intervals $\{[j_1, \ldots, j_1 + \ell_1 - 1], \ldots, [j_m, \ldots, j_m + \ell_m - 1]\}$ having length $\ell_1, \ldots, \ell_m$, is a* generalized covering *of the interval $\mathcal{I}$ if the following conditions hold:*

- *The intersection of every interval in the set with $\mathcal{I}$ is non-empty.*
- *Each interval in the set is either an interval of length 1 or the RPOI of a phrase $y \in \mathbf{y}$ for which $I(y) \cap \mathcal{I} = \emptyset$.*
- *The union of the intervals in the set contains $\mathcal{I}$.*

The cardinality of $\{[j_1, \ldots, j_1 + \ell_1 - 1], \ldots, [j_m, \ldots, j_m + \ell_m - 1]\}$ is m, the number of intervals.

From the definition it follows that every covering is also a generalized covering. Among all generalized coverings of an interval, we are interested in the ones with the smallest cardinality.

Definition 7 *A minimal generalized covering $\mathcal{C}_i^k(j)$ of the interval $\mathcal{I}$ at time j is a generalized covering with smallest cardinality among those whose elements are intervals of length 1 or RPOIs of phrases ending at or before position j.*

In general the minimal generalized covering of $\mathcal{I}$ at time j is not unique. As we parse $\mathbf{x}$ from left to right, we keep track of the minimal generalized coverings that can be obtained using already parsed phrases.

Definition 8 *The cardinality of a minimal generalized covering $\mathcal{C}_i^k(j)$ is denoted by $\mathcal{N}_i^k(j)$.*

It is immediate from the definition that $\mathcal{N}_i^k(l) \leq k$ because the k intervals of length 1 starting at $i, i+1, \ldots, i+k-1$ form a covering of $\mathcal{I}$. Also, trivially, $\mathcal{N}_i^k(l) \geq 1$. The cardinality of the minimal generalized coverings of $\mathcal{I}$ is a non-increasing function of the current position j. In the following, we will be interested in the changes in the size of the minimal generalized coverings, and therefore we introduce the following notation.

Definition 9 *Let*

$$\Delta\mathcal{N}_i^k(j) \stackrel{\triangle}{=} \mathcal{N}_i^k(j-1) - \mathcal{N}_i^k(j).$$

It is immediate to check the following statements.

Observation 1 *The quantity $\Delta\mathcal{N}_i^k(j)$ is non-negative, has maximum value equal to $k-1$, and $\sum_j \Delta\mathcal{N}_i^k(j) \leq k-1$.*

We are now ready to state the main result of this section.

Theorem 4 *Let* $\underline{L} = \max\left\{\mathbb{L} \mid \sum_{\ell=2}^{\mathbb{L}} \ell\Big(\min(i,\ell)+k-1\Big) \leq n-i-k-\mathbb{L}+2\right\}$ *and* $\overline{L} = \min\left\{\ell \mid \ell^3 + \frac{3}{2}k\ell^2 + \frac{3k-2}{2}\ell \geq 3(n-k-i-1)\right\}$. *The maximum expansion of a string when the k consecutive symbols located in the interval $\mathcal{I}$ starting from position i are changed, satisfies*

$$\max_{\mathbf{y},\hat{\mathbf{y}}[i]} (\|\hat{\mathbf{y}}\| - \|\mathbf{y}\|) \leq \frac{\overline{L}\left(\overline{L}+1\right)}{2} + \left\lceil \left(\sqrt{\frac{(k+1)^2}{4} + (n-i-k+1)} - \frac{(k+1)}{2}\right)\right\rceil + k\log_2 k + 5k + 2 = \overline{\Delta}_i^k, \quad (15)$$

and, if the alphabet size is at least $\min(i,\underline{L}) + k + \underline{L}$,

$$\underline{\Delta}_i^k = \sum_{\ell=2}^{\underline{L}} \min(i,\ell) + k - 1 \quad \leq \quad \max_{\mathbf{y},\hat{\mathbf{y}}[i]} (\|\hat{\mathbf{y}}\| - \|\mathbf{y}\|). \quad (16)$$

As before, we first provide an overview of the proof, we state and prove the supporting lemmas for the upper bound, we prove the upper bound, and finally prove the lower bound.

A simple extension of Lemma 2 to the case of k consecutive changes shows that we need not concern ourselves with phrases of $\mathbf{y}$ and $\hat{\mathbf{y}}$ starting and ending before $\mathcal{I}$.

Hence, we first consider phrases overlapping $\mathcal{I}$ and extend Lemma 1 to $k \geq 1$. The extension lemma, Lemma 11, states that the cumulative contribution of these phrases is at most $2k+2$ (e.g., at most 4 if $k=1$, as expected).

We then consider phrases pointing to but not overlapping $\mathcal{I}$. When $k > 1$ consecutive symbols are changed, the fundamental result of Lemma 4, which states that a phrase pointing to but does not overlap the changed position i has $\phi(\cdot) \leq 2$, is no longer valid. It is extended by Lemma 13 which states that only phrases that reduce the size of the minimal generalized covering(s) can have $\phi(\cdot) > 2$, and that, for any such phrase y, the quantity $(\phi(y) - 2)$ is at most equal to the reduction in size of the minimal generalized covering. This result, combined with Observation 1, allows us to bound from above the contribution of phrases not pointing to $\mathcal{I}$ to the sum in Inequality 1 by counting the number of phrases that could point to $\mathcal{I}$, adding a bound to the number of phrases that could have $\phi(\cdot) > 1$, and further adding $(k-1)$ to the result, to account for the cumulative contribution of phrases with $\phi(\cdot) > 2$. To bound the maximum number of phrases that could point to $\mathcal{I}$ we use an approach analogous to that used in proving Theorem 2. We need to exercise some care in bounding the number of phrases with $\phi(\cdot) \geq 2$. To this end, we divide these phrases into three groups according to their RPOI, and analyze each group separately. More specifically, we consider the phrases with RPOI entirely contained in $\mathcal{I}$, the phrases with RPOI strictly containing $\mathcal{I}$, and the remaining phrases, whose RPOI is not contained in $\mathcal{I}$ and partially overlaps it.

The lower bound is again proved by bounding from below the maximum number of phrases that point to $\mathcal{I}$ and constructing a string satisfying this bound, where each phrase has $\phi(\cdot)$ exactly equal to 1.

5.1 The Upper Bound of Theorem 4

We address the question of the maximum contribution to the sum in Inequality 1 due to phrases that overlap the interval $\mathcal{I}$. The following lemma provides the desired extension to Lemma 1 to the case of $k \geq 1$ consecutive changes.

Lemma 11 *Let $\mathcal{Y} \stackrel{\triangle}{=} \{y \in \mathbf{y} \mid I(y) \bigcap \mathcal{I} \neq \emptyset\}$ be the set of phrases overlapping the interval $\mathcal{I}$. Then,*

$$\sum_{y \in \mathcal{Y}} \phi(y) \leq 2k + 2. \tag{17}$$

Note that the bound is equal to 4 when $k = 1$, and therefore in this case Lemma 11 reduces to Lemma 1.

Proof. We show that there are at most $2k + 3$ phrases in $\hat{\mathbf{y}}$ contained in $I(\mathcal{Y}) \stackrel{\triangle}{=} \bigcup_{y \in \mathcal{Y}} I(y)$, that if there are exactly $2k + 3$ phrases the last phrase ends at the end of $I(\mathcal{Y})$, that $\|\mathcal{Y}\| \geq 1$, and

therefore

$$\sum_{y\in\mathcal{Y}} \phi(y) \leq 2k+3-1 = 2k+2.$$

The "-1" term is due to the fact that if there are $2k+3$ phrases of $\hat{\mathbf{y}}$ contained in $I(\mathcal{Y})$, the last one ends at $\max\{I(\mathcal{Y})\}$.

To simplify the notation, renumber the phrases of $\mathbf{y}$ so that $\mathcal{Y} = \{y_1, \ldots, y_p\}$. Phrases to the left of $I(\mathcal{Y})$ will therefore have index less than or equal to zero, say, from $-q$ to 0. Since the compression schemes we consider in this paper only allow pointers to the past (and not to the future), clearly $y_{-q} = \hat{y}_{-q}, \ldots, y_{-1} = \hat{y}_{-1}$. Also, either $y_0 = \hat{y}_0$, or $|y_0| < |\hat{y}_0|$, because the symbols in the RPOI of y_0 are the same in both $\mathbf{x}$ and $\hat{\mathbf{x}}$.

First, there are at most k phrases of $\hat{\mathbf{y}}$ overlapping $\mathcal{I}$. Among the phrases in $\mathcal{Y}$ there is at most one, y_1, that starts before position i, and at most one, y_p, that ends after position $i+k-1$. If y_1 starts before position i, say, at position $i-a$, then $\hat{y}_1$ (if it exists) ends at or after position $i-1$, because $x_{i-a}^{i-1} = \hat{x}_{i-a}^{i-1}$ and $\mathrm{RPO}(x_{i-a}^{i-1})$ is the same in $\mathbf{x}$ and $\hat{\mathbf{x}}$. Then, let $I(y_p) = [b, \ldots, c]$, and consider phrases of $\hat{\mathbf{y}}$ starting at or after position $i+k$ and ending before or at the end of y_p. The $\mathrm{RPOI}(y_p)$ overlaps $\mathcal{I}$ in at most k positions, and therefore there are at most k phrases in $\hat{\mathbf{y}}$ overlapping the interval containing the symbols of y_p that were copied from anywhere within $\mathcal{I}$. Let these symbols be x_d^e. If $d > i+k$, then there could be at most one phrase of $\hat{\mathbf{y}}$ starting at or after position $i+k$ and ending before position d, because $x_{i+k}^{d-1} = x_{2i-d+k}^{i-1} = \hat{x}_{2i-d+k}^{i-1} = \hat{x}_{i+k}^{d-1}$. If $c > e$ then there is at most one phrase of $\hat{\mathbf{y}}$ starting at or after position $e+1$ and ending at or after c, because $x_{e+1}^{c} = \hat{x}_{e+1}^{c} = x_{i+k}^{i+k+c-e-1} = \hat{x}_{i+k}^{i+k+c-e-1}$. If such phrase exists, then it must end at or after $\max\{I(\mathcal{Y})\}$.

We conclude that there are at most $2k+3$ phrases of $\hat{\mathbf{y}}$ contained in $I(\mathcal{Y})$, and if there are exactly $2k+3$ such phrases, the last one ends at $\max\{I(\mathcal{Y})\}$. □

It is worth noting that for k greater than 1, there are always fewer than $2k+3$ phrases of $\hat{\mathbf{y}}$ contained in $I(\mathcal{Y})$, and that therefore the bound can be further tightened. We now bound from above the maximum number of phrases that can point to $\mathcal{I}$ without overlapping it, and hence extend Lemma 3 to the case of k consecutive changes.

Lemma 12 *The maximum number N_5 of phrases $y \in \mathbf{y}$ such that $I(y) \bigcap \mathcal{I} = \emptyset$ and $RPOI(y) \bigcap \mathcal{I} \neq \emptyset$ is bounded from above by*

$$N_5 \leq \frac{\overline{L}\left(\overline{L}+1\right)}{2}, \tag{18}$$

where $\overline{L} = \min\left\{\ell \mid \ell^3 + \frac{3}{2}k\ell^2 + \frac{\ell}{2}(3k-2) \geq 3(n-i-k+1)\right\}$.

Note that Inequality 18 is formally identical to Inequality 5 and that these results differ in the definition of $\overline{L}$.

Proof. There are at most $k+\ell-1$ phrases of length ℓ whose RPOI intersects the interval $\mathcal{I}$. As in the proof of Lemma 3, construct a pseudo-parsing by appending, to the right of $\mathcal{I}$, k pseudo-phrases of length one, $k+1$ pseudo-phrases of length two and so on until the $n-i-k+1$ positions to the right of $\mathcal{I}$ are filled. For no string of length n the LZ'77 parsing can produce more phrases that point to the interval $\mathcal{I}$ and start at or after position $i+k$ than this pseudo-parsing, because the pseudo-parsing uses all the available short pseudo-phrases.

The number of positions occupied by pseudo-phrases of length $\leq \ell$ is

$$\begin{aligned}\sum_{j=1}^{\ell}(k+j-1)*j &= (k-1)\sum_{j=1}^{\ell} j + \sum_{j=1}^{\ell} j^2 \\ &= (k-1)\frac{\ell(\ell+1)}{2} + \frac{\ell(\ell+1)(2\ell+1)}{6} \\ &= \frac{1}{3}\left(\ell^3 + \frac{3}{2}k\ell^2 + \frac{3k-2}{2}.\ell\right)\end{aligned}$$

Hence, the longest pseudo-phrase contains at most

$$\overline{L} = \min\left\{\ell \mid \sum_{j=1}^{\ell} j(k+j-1) \geq (n-i-k+1)\right\}$$

symbols, or, equivalently,

$$\overline{L} = \min\left\{\ell \mid \ell^3 + \frac{3}{2}k\ell^2 + \frac{3k-2}{2}\ell \geq 3(n-i-k+1)\right\}.$$

□

We now analyze phrases that point to $\mathcal{I}$ without intersecting it, and with $\phi(\cdot) > 1$. We first show that an upper bound to their cumulative contribution to the sum in Inequality 1 can be obtained by adding $k-1$ to an upper bound to the number of such phrases multiplied times 2.

To bound from above the number of phrases with $\phi(\cdot) > 1$, we divide them according to their RPOI. First we consider phrases with RPOI completely contained in $\mathcal{I}$ (including a possible phrase whose RPOI coincides with $\mathcal{I}$). We then analyze phrases whose RPOI strictly contains $\mathcal{I}$ (i.e., with padding at least at one end), and finally consider the remaining phrases.

Lemma 13 *Let y be a phrase of $\mathbf{y}$ pointing to but not overlapping the interval $\mathcal{I}$, and ending at position j. Then*

$$\phi(y) \leq 2 + \Delta\mathcal{N}_i^k(j).$$

Proof. Consider first the case $\Delta\mathcal{N}_i^k(j) = 0$. Then, there exists at least one minimal generalized covering at time $j-1$, say $\mathcal{C}^{*k}_i(j-1)$ that is also a minimal generalized covering at time j, since $\mathcal{N}_i^k(j-1) = \mathcal{N}_i^k(j)$. $\mathcal{C}^{*k}_i(j-1)$ does not contain RPOI(y), which means that there is at most one interval of $\mathcal{C}^{*k}_i(j-1)$ completely contained in RPOI(y) (otherwise we could construct a new minimal generalized covering, using RPOI(y) and the intervals of $\mathcal{C}^{*k}_i(j-1)$ not completely contained in RPOI(y), that has cardinality smaller than $\mathcal{N}_i^k(j)$, which is a contradiction).

Consider now the case $\Delta\mathcal{N}_i^k(j) = \delta > 0$. Every minimal generalized covering at time j must contain RPOI(y), because it is the only interval not available to construct generalized minimal coverings at time $j-1$. The interval RPOI(y) can contain at most $\delta+1$ intervals of any minimal generalized covering at time $j-1$. By contradiction: assume there exists a minimal generalized covering at time $j-i$, say $\mathcal{C}^{*k}_i(j-1)$, containing $\delta+1+m$ intervals ($m>0$) that are subintervals of RPOI(y). Then the union of RPOI(y) and of the intervals of $\mathcal{C}^{*k}_i(j-1)$ not completely contained in RPOI(y) form a generalized covering having cardinality smaller than $\mathcal{N}_i^k(j)$.

Hence, RPOI(y) completely contains at most $\Delta\mathcal{N}_i^k(j)+1$ intervals of any minimal generalized covering at time $j-1$.

Parse the symbols of $\hat{\mathbf{x}}$ in $I(y)$ from left to right. Our usual arguments show that $\phi(y)$ cannot be larger than one plus the maximum number of intervals of any minimal generalized covering at time $j-1$ completely contained in RPOI(y), namely, $\phi(y) \leq \Delta\mathcal{N}_i^k(j)+2$. □

Lemma 14 *The number of phrases with RPOI contained in the interval $\mathcal{I}$, that start after $i+k-1$ and with a value of ϕ greater than 1 is at most $k\log_2 k$.*

Proof. We say that a phrase y is of interest if $I(y) \geq i+k$, $\phi(y) > 1$ and RPOI(y) $\subset \mathcal{I}$. Let $J = [r, \cdots, r+l-1]$ be any interval with positive length l that is contained in the interval $\mathcal{I} = [i, \cdots, i+k-1]$. Now assume that $l \geq 3$. If l is odd, we define the pivot P_J to be the interval whose only member is the integer $r+(l-1)/2$ and, if l is even, the interval $[r+l/2-1, r+l/2]$. Also, for l odd (resp. l even), let J_0 be the interval $[r, \ldots, r+(l-3)/2]$ (resp. $[r, \ldots, r+l/2-2]$) and let J_1 be the interval $[r+(l+1)/2, \ldots, r+l-1]$ (resp. $[r+l/2+1, \ldots, r+l-1]$). The procedure that is described next can be carried out for any string and choice of potentially modified symbols in $\mathcal{I}$.

We assign every phrase of interest to one of three categories: phrases for which $P_J \bigcap \text{RPOI}(y)$ is nonempty (we say that the RPOIs of these phrases straddles the pivot), phrases with $\text{RPOI}(y) \subset J_0$, and phrases with $\text{RPOI}(y) \subset J_1$. Note that the RPOI of phrases that straddle the pivot need not contain the pivot itself: we require non-empty intersection in order to count all the phrases of interest with RPOI included in J. If their length is at least three, each of J_0 and J_1 also possess a pivot and can be partitioned similarly; here we will use the convention that sub-indexes can be appended to the right of existing ones so that for example $(J_0)_0 = J_{00}$ and $(J_1)_0 = J_{10}$. This notation extends obviously to finer partitioning.

The number of phrases of interest with RPOI contained in J will be denoted by $\#(J)$. The number of phrases that in addition of satisfying these conditions also straddle the pivot of J will be denoted by $\#_P(J)$. Thus,

$$\#(J) = \#_P(J) + \#(J_0) + \#(J_1). \tag{19}$$

Note that all the quantities in this expression depend on a particular choice of string and associated modifications. Nevertheless, we shall shortly prove that

$$\#_P(J) \leq |J|. \tag{20}$$

It can be verified that $|J_0| \leq \lfloor |J|/2 \rfloor$, an identical statement obviously being true for the cardinality of J_1. Substituting $\mathcal{I}$ for J in (19), repeating the same procedure for the intervals $\mathcal{I}_0$ and $\mathcal{I}_1$ and using (20) we get

$$\begin{aligned} \#(\mathcal{I}) &\leq k + 2\lfloor k/2 \rfloor + \#(\mathcal{I}_{00}) + \#(\mathcal{I}_{01}) + \#(\mathcal{I}_{10}) + \#(\mathcal{I}_{11}) && (21) \\ &\leq 2k + \#(\mathcal{I}_{00}) + \#(\mathcal{I}_{01}) + \#(\mathcal{I}_{10}) + \#(\mathcal{I}_{11}) && (22) \end{aligned}$$

Repeating this procedure r times when possible (the above corresponding to $r = 2$), we obtain

$$\#(\mathcal{I}) \leq rk + \sum_{\text{mask} \in \{0,1\}^r} \#(\mathcal{I}_{\text{mask}}) \tag{23}$$

By construction, each of the intervals $\mathcal{I}_{\text{mask}}$ in the summation above has the same length for a fixed value of r. This process cannot be repeated indefinitely for the length of said intervals strictly decreases as r increases and pivots are defined only for intervals of length at least three. Let r^* be the largest value of r for which we can write the associated Inequality (23); the length of every

interval in the right hand side is 1 or 2. Since every phrase that has $\phi(\cdot) > 1$ must have length at least three,

$$\#(\mathcal{I}) \leq r^* k. \tag{24}$$

On the other hand, it is easy to see that if mask $\in \{0,1\}^{r^*}$, the strict inequality

$$1 \leq |\mathcal{I}_{\text{mask}}| < k/2^{r^*}$$

holds and therefore $r^* < \log_2 k$. In combination with (24), this yields the conclusion of the lemma.

It remains to prove that the number of phrases of interest with RPOI completely contained in an interval J that straddle J's pivot is at most $|J|$ (Inequality (20)), irrespective of the choice of string and associated modified symbols. For any $i + k - 1 < j \leq n + 1$, let

$$S(j) = \left\{ y : y \text{ is of interest, } P_J \bigcap \text{RPOI}(y) \text{ is nonempty, } \text{RPOI}(y) \subset J \text{ , and } i + k - 1 < I(y) < j \right\}.$$

Rephrasing our goal using this notation, we want to demonstrate that $|S(n+1)| \leq |J|$. Let Γ_0^j (resp. Γ_1^j) be the smallest (resp. largest) integer in the interval

$$\bigcup_{y \in S(j)} \text{RPOI}(y).$$

Now suppose that in the original parsing, a phrase of interest y of length ℓ has been copied at position j, and that $y \in S(j + \ell)$. We next show that if $\text{RPOI}(y) \subset [\Gamma_0^j, \ldots, \Gamma_1^j]$ then necessarily $\phi(y) \leq 1$, regardless of whether $|J|$ is even or odd. The argument is as follows: as a basic consequence of the definitions, there exists a set of phrases $\Pi \subset S(j)$ with cardinality $|\Pi| = 2$ and with the property that

$$\bigcup_{\pi \in \Pi} \text{RPOI}(\pi) = [\Gamma_0^j, \ldots, \Gamma_1^j]. \tag{25}$$

Note that it is not possible that $|\Pi| = 1$, as otherwise $\text{RPOI}(y) \subset I(\pi)$, assuming π is the single member of Π (recall that the parsing rule always selects the *rightmost* match for a given maximum match length). It is easy to see that it is possible to segment the phrase y in at most two pieces, each of which is a match to a subset of a phrase $\pi \in \Pi$. Since the symbols corresponding to phrases in Π remain unchanged in the modified string, necessarily $\phi(y) \leq 1$.

Since we have deduced that strict containment $\text{RPOI}(y) \subset [\Gamma_0^j, \ldots, \Gamma_1^j]$ is not possible, a fortiori,

$$\Gamma_1^{j+\ell} - \Gamma_0^{j+\ell} \geq \Gamma_1^j - \Gamma_0^j + 1.$$

Clearly, $|S(i+k)| = 0$ and it is always true that for any i, $\Gamma_1^i - \Gamma_0^i \leq |J|$. Note that for any $j \geq i+k$, $|S(j+1)| - |S(j)| \in \{0,1\}$ and, as per our earlier discussion, the number of indexes j where $|S(j+1)| - |S(j)| = 1$ cannot be greater than $|J|$. Therefore,

$$|S(n+1)| = \sum_{j=i+k}^{n} |S(j+1)| - |S(j)| \leq |J|$$

and the lemma is proved. □

Lemma 15 *The number of phrases not overlapping $\mathcal{I}$, with RPOI containing $\mathcal{I}$ with padding at both ends, and with $\phi(\cdot)$ greater than 1 is bounded from above by*

$$N_4 = \min\left(N_4^{'}, i-1\right),$$

where

$$N_4^{'} = \min\left\{ m \mid \sum_{\ell=1}^{m} k + 2\ell \geq n-i-k+1 \right\} \leq \left\lceil \sqrt{\frac{(k+1)^2}{4} + (n-i-k+1)} - \frac{(k+1)}{2} \right\rceil$$

Proof. Using Lemma 5, we see immediately that two phrases with RPOI containing the interval $\mathcal{I}$, not overlapping $\mathcal{I}$, and with $\phi(\cdot) > 1$ must have lengths that differ at least by 2. These phrases also must have length at least equal to $k+2$. If m phrases satisfy the conditions of the lemma, they occupy at least $mk + 2m(m+1)/2 = m^2 + m(k+1)$ positions. There are $n-i-k+1$ positions after $\mathcal{I}$, hence the maximum number $N_4^{'}$ of such phrases is bounded from above by the smallest integer larger than or equal to the positive solution of $m^2 + m(k+1) = n-i-k+1$.
The solution of this equation is $m^* = \left(\sqrt{(k+1)^2 + 4(n-i-k+1)} - (k+1)\right)/2$, which is
$\leq \left\lceil \left(\sqrt{(k+1)^2 + 4(n-i-k+1)} - (k+1)\right)/2 \right\rceil \stackrel{\triangle}{=} N_4^{'}$. Lemma 5 ensures that two phrases whose RPOIs contain the interval $[i, \ldots, i+k-1]$ and having both $\phi(\cdot) > 1$ are such that one contains the other with padding at both ends. Hence, there can be at most $i-1$ such phrases. □

The following lemma bounds the number of phrases that do not intersect the interval $\mathcal{I}$, with RPOI that overlaps $\mathcal{I}$ with padding at one end only, and with $\phi(\cdot) > 1$.

Lemma 16 *The number N_1 of phrases y not intersecting the interval $\mathcal{I}$, with RPOI starting before position i and ending before or at position $i+k-1$, and with $\phi(y) > 1$ is at most equal to k.*
The number N_2 of phrases y not intersecting the interval $\mathcal{I}$, with RPOI starting at or after position i and ending after position $i+k-1$, for which $\phi(y) > 1$ is at most equal to k.

Therefore the total number of phrases that do not overlap $\mathcal{I}$, having RPOI that overlap $\mathcal{I}$ without containing it and with padding at one end only, and having $\phi(\cdot)$ greater than 1 is at most $2k$.

Proof. We only prove the first part of the lemma, the second part being analogous. Let y_1 be a phrase with $I(y_1) = [p-a, \ldots, p+b]$ and RPOI $= [i-a, \ldots, i+b]$, with $a > 0$ and $0 \leq b \leq k-1$. Let y_2 be a subsequent phrase with $I(y_2) = [q-c, \ldots, q+d]$ and RPOI $= [i-c, \ldots, i+d]$, with $c > 0$ and $0 \leq d \leq k-1$. The RPOIs of both phrases intersect $\mathcal{I}$ with padding at the left end only.

We claim that $d \leq b$ implies $\phi(y_2) \leq 1$. To see this, note first that x_q^{q+d} is equal to x_i^{i+d}, which in turn is equal to x_p^{p+d} by assumption. Since neither y_1 nor y_2 overlap the interval $\mathcal{I}$, then $\hat{x}_q^{q+d} = \hat{x}_p^{p+d}$. Change the symbols in positions $\mathcal{I}$, parse $\hat{\mathbf{x}}$ to obtain $\hat{\mathbf{y}}$, and restrict the attention to the interval $I(y_2) = [q-c, \ldots, q+d]$. The parsing $\hat{\mathbf{y}}$ contains a phrase $\hat{y}$ starting at position $e \leq q-c$ and ending at position $f \geq q-c$.

If $f \geq q+d$, then $I(y_2) \subseteq I(\hat{y})$, and $\phi(y_2) = 0$. Divide the case where $f < q+d$ into two sub-cases: $f \geq q-1$ and $f < q-1$ (where perforce $c > 1$). If $f \geq q-1$, then the phrase following $\hat{y}$ in $\hat{\mathbf{y}}$ cannot end before position $q+d$, because $\hat{x}_{q-1}^{q+d} = \hat{x}_{p-1}^{p+d}$ and $p < q$. Therefore $\phi(y_2) \leq 1$. If $f < q-1$, then $e < q-c$ (because $\hat{x}_{p-c}^{p-1} = \hat{x}_{q-c}^{q-1}$) and the phrase $\hat{y}'$ of $\hat{\mathbf{y}}$ that follows $\hat{y}$ must end at or after position $q-1$ because $\hat{x}_{q-c}^{q-1} = \hat{x}_{i-c}^{i-1}$; the next phrase of $\hat{\mathbf{y}}$, $\hat{y}''$ starts at or after q and cannot end before $q+d$ because $\hat{x}_q^{q+d} = \hat{x}_p^{p+d}$. We conclude that $\phi(y_2) \leq 1$ in this case too.

Hence, $\phi(y_2)$ can be greater than 1 only if $d > b$. This can happen only k times for the set of sequences satisfying the conditions of the lemma. □

Proof of the Upper Bound

To prove the upper bound, we note that phrases with $\phi(\cdot) \geq 1$ overlap or point to the interval $\mathcal{I}$ where $\mathbf{x}$ and $\hat{\mathbf{x}}$ differ, as stated in Corollary 2. We consider separately the phrases that point to and do not overlap $\mathcal{I}$, and the phrases that overlap $\mathcal{I}$.

As in the proof of Theorem 2, we count the maximum number of phrases that can point to $\mathcal{I}$ without overlapping it: Lemma 12 provides the necessary bound. Then we bound the additional contribution of phrases that point to and do not overlap $\mathcal{I}$, and that have $\phi(\cdot) > 1$: Lemma 13 states that if y points to $\mathcal{I}$ and does not overlap $\mathcal{I}$, $\phi(y) \leq 2 + \Delta\mathcal{N}_i^k(j)$, where j is the ending position of y. By definition, $\sum \Delta\mathcal{N}_i^k(j) \leq k-1$, where the sum is over all phrases pointing to $\mathcal{I}$. Hence, the additional contribution of phrases pointing to and not overlapping $\mathcal{I}$ and with $\phi(\cdot) > 1$ is bounded by the number of such phrases plus $k-1$. To compute a bound on the number of phrases

not overlapping $\mathcal{I}$, pointing to $\mathcal{I}$, and having $\phi(\cdot) > 1$, we divide them into four sets: phrases with RPOI starting before position i and ending before or at position $i + k - 1$, phrases with RPOI starting at or after position i and ending at or after position $i + k$ (whose number is bounded from above in Lemma 16), phrases with RPOI contained in $\mathcal{I}$ (whose number is bounded from above in Lemma 14), and phrases with RPOI containing $\mathcal{I}$ with padding at both ends (whose number is bounded from above in Lemma 15).

Therefore, the upper bound (15) is obtained as the sum of the bound on the contribution of phrases overlapping $\mathcal{I}$ (Equation (17) in Lemma 11) the bound on the number of phrases pointing to and not overlapping $\mathcal{I}$ (Equation (18) in Lemma 12), and the bound on the additional contribution of phrases pointing to and not overlapping $\mathcal{I}$ and having $\phi(\cdot) > 1$. □

5.2 The Lower Bound of Theorem 4

The following lemma provides a lower bound to the maximum number of phrases that can point to the interval $\mathcal{I}$ in the parsing $\mathbf{y}$ of a string $\mathbf{x}$ having length n, and is the equivalent of Lemma 8.

Lemma 17 *Let*

$$\underline{L} = \max\left\{\mathbb{L} \mid \sum_{\ell=2}^{\mathbb{L}} \ell\Big(\min(i,\ell) + k - 1\Big) \leq n - i - k - \mathbb{L} + 2\right\}.$$

If the alphabet size is at least $\min(i,\underline{L}) + k + \underline{L}$*, the maximum number* N_0 *of phrases of* $\mathbf{y}$ *pointing to* $\mathcal{I}$ *is bounded from below by*

$$N_0 \geq \sum_{\ell=2}^{\underline{L}} \min(i,\ell) + k - 1. \tag{26}$$

Proof. We construct a string $\mathbf{x}$ that satisfies Inequality 26, using the following packing algorithm. Let $x_{i-\min(i,\underline{L})+1}^{i+k+\underline{L}-2}$ be distinct symbols of the alphabet. There is at least one unused alphabet symbol, which we denote by $\tilde{\alpha}$.

We now consider only intervals overlapping $\mathcal{I}$ with length not exceeding $n - i - k + 2$. It is easy to see that there are $\min(i,\ell) + k - 1$ such intervals of length $\ell \leq n - i - k + 2$. Now order all the intervals of length 2 so that the first interval in the list is the rightmost one and the last interval is the leftmost one. Next, copy the substrings corresponding to these intervals as follows: the first substring is copied into positions $i + \underline{L} + k - 1$ and $i + \underline{L} + k$, the second substring is copied immediately after that and so on until the last substring on the list is copied.

Next the intervals of length 3 are ordered in the opposite direction, so that for example the first interval in the list is now the leftmost one. The substrings corresponding to the intervals are now copied following the order in the list as above. This process is repeated until all the substrings of length $\underline{L}$ overlapping $\mathcal{I}$ have been copied, always reversing the direction in which the list is ordered when increasing the substring length. The definition of $\underline{L}$ ensures that there is sufficient space in $\mathbf{x}$ to contain all these copies. If there is additional space at the beginning or at the end of $\mathbf{x}$, fill it with copies of $\tilde{\alpha}$. Parse $\mathbf{x}$ into $\mathbf{y}$. The first $i-1$ symbols are irrelevant, and the symbols in the interval $[i,\ldots,i+k+\underline{L}-2]$ parse into phrases of length 1. Note that the last symbol of each copied substring and the first symbol of the next copied substring appear only once as a contiguous pair in the entire $\mathbf{x}$. Hence, when parsing $\mathbf{x}$ these symbols are always separated by a comma. Also, each copied substring by construction has a non-empty RPOI that intersects $\mathcal{I}$. Hence, $\mathbf{y}$ contains exactly $\sum_{\ell=1}^{\underline{L}} \min(i,\ell)+k-1$ phrases pointing to the interval $\mathcal{I}$. □

Proof of the Lower Bound

As for Theorem 2, the lower bound is constructive. Lemma 17 constructively shows that the maximum number N_0 of phrases pointing to $\mathcal{I}$ is bounded from below by $\sum_{\ell=2}^{\underline{L}} \min(\ell,i)+k-1$, if the alphabet size is sufficiently large, say $\geq \min(i,\underline{L})+k+\underline{L}$. We use the string $\mathbf{x}$ and the "unused symbol" $\tilde{\alpha}$ defined in Lemma 17. We build $\hat{\mathbf{x}}$ by changing the symbols of $\mathbf{x}$ in $\mathcal{I}$ to $\tilde{\alpha}$. The string $\mathbf{x}$ is carefully crafted to ensure that each pair of symbols straddling commas in the parsing (i.e., the last symbol of a phrase and the first symbol of the following phrase) appear only once in both $\mathbf{x}$ and $\hat{\mathbf{x}}$. Hence, each comma of $\mathbf{y}$ has a corresponding comma in $\hat{\mathbf{y}}$, and for each phrase y in $\mathbf{y}$ there is a phrase $\hat{y}(y)\in\Phi(y)$ in $\hat{\mathbf{y}}$ starting at $\min I(y)$. Additionally, $|y|>1$ implies $|\hat{y}(y)|<|y|$, because y has a unique PO and $\mathcal{I}\bigcap \mathrm{RPOI}(y)\neq\emptyset$. We now show that $\mathrm{PO}(y)$ is unique. The symbol $\tilde{\alpha}$ does not appear in phrases pointing to $\mathcal{I}$. The phrase y has a unique PO intersecting the interval $[i-\underline{L}+1,\ldots,i+k+\underline{L}-2]$ because the symbols in this interval are different. No phrase to the right of position $i+k+\underline{L}-2$ and to the left of y is equal to y by construction. It remains to be shown that y cannot be copied from any other substring of $\mathbf{x}$ starting to the to the right of position $i+k+\underline{L}-2$ and to the left of y, and perforce intersecting at least two adjacent phrases having a PO that contains $\mathcal{I}$ (where the rightmost of such phrases could be y itself). This is not possible, because for each pair of adjacent phrases each having a PO that intersects $\mathcal{I}$, the last symbol of the first phrase and the first symbol of the second phrase appear as a pair only once

in the entire string $\mathbf{x}$, as discussed in the proof of Lemma 17. Hence, each phrase $y \in \mathbf{y}$ with $\mathcal{I} \bigcap \mathrm{RPOI}(y) \neq \emptyset$ has $\phi(y) \geq 1$. But $\phi(y) \leq 1$ because each phrase $y \in \mathbf{y}$ with $\mathcal{I} \bigcap \mathrm{RPOI}(y) \neq \emptyset$ and $|y| > 2$ is preceded by a phrase $y^{'}(y) \in \mathbf{y}$ satisfying $\left|y^{'}(y)\right| \geq |y| - 1$, $\mathcal{I} \bigcap \mathrm{RPOI}(y^{'}) \neq \emptyset$, and $\left|\mathrm{RPOI}(y) \bigcap \mathrm{RPOI}(y^{'})\right| = |y| - 1$. We conclude that each phrase $y \in \mathbf{y}$ with $|y| \geq 2$ has $\phi(y)$ exactly equal to 1. □

6 Result Refinements

We now refine the upper bound of Theorem 4 using arguments analogous to those of Theorem 3. To this end, we first introduce a "substitution" lemma, namely, an extension of Lemma 9 to the case of k consecutive changes, that will allow us to extend the pairing argument of Theorem 3 to $k > 1$ consecutive changes.

Lemma 18 *Consider a phrase y with $I(y) = [m - a, \ldots, m + k + b - 1]$, $\phi(y) = d > 1$, and RPOI(y)$= [i - a, \ldots, i + k - 1 + b]$.*

- *If $a > b$, for every $j < m - a$ such that $[j, \ldots, j + k + a - 1] \bigcup \mathcal{I} = \emptyset$, $x_j^{j+k+a-1} \neq x_{i-a}^{i+k-1}$, and, for every $j < m - a$ such that $[j, \ldots, j + k + a] \bigcup \mathcal{I} = \emptyset$, $x_j^{j+k+a} \neq x_{i-a}^{i+k}$.*
- *If $a < b$, for every $j < m$ such that $[j, \ldots, j + k + b - 1] \bigcup \mathcal{I} = \emptyset$, $x_j^{j+k+b-1} \neq x_{i-1}^{i+k+b-1}$, and, for every $j < m - a$ such that $[j, \ldots, j + k + b] \bigcup \mathcal{I} = \emptyset$, $x_j^{j+k+b} \neq x_i^{i+k+b}$.*
- *If $a = b$, for every $j < m - a$ such that $[j, \ldots, j + k + a - 1] \bigcup \mathcal{I} = \emptyset$, $x_j^{j+k+a-1} \neq x_{i-a}^{i+k-1}$ and, for every $j < m$ such that $[j, \ldots, j + k + b - 1] \bigcup \mathcal{I} = \emptyset$, $x_j^{j+k+b-1} \neq x_i^{i+k+b-1}$.*

The proof is analogous to that of Lemma 9 and is left to the interested reader.

Theorem 5 *Under the conditions of Theorem 4,*

$$\max_{\mathbf{y}, \hat{\mathbf{y}}[i]} \left(\|\hat{\mathbf{y}}\| - \|\mathbf{y}\|\right) \leq \frac{M(M+1)}{2} + k \log_2 k + 4k + 1, \tag{27}$$

where

$$\begin{aligned} M \;=\; & \min \Bigg\{ \ell \mid \ell^3 + \frac{3}{2} k \ell^2 + \frac{\ell}{2}(3k - 2) \\ & \geq 3 \left[n - k - i + k \left[\sqrt{\frac{(k+1)^2}{4} + (n - i - k + 1)} - \frac{(k+1)}{2} \right] \right] \Bigg\}. \end{aligned}$$

Proof. First, we note that the term $5k$ in Theorem 4 can be immediately reduced to $4k$, because Lemmas 16 and 11 interact and over-count contributions to the sum of $\phi(\cdot)$. Consider in fact the last phrase y_p in the proof of Lemma 11: this phrase ends after position $i+k-1$. Consider the portion of y_p that does not overlap $\mathcal{I}$, and count the number $h \leq k$ of symbols of this portion that are copied from $\mathcal{I}$ in the parsing $\mathbf{y}$ of $\mathbf{x}$. Consider then the phrases of $\mathbf{y}$ analyzed in Lemma 16: they do not intersect $\mathcal{I}$, and therefore start after y_p and have RPOI that starts before position i and ends before or at position $i+k-1$. Since h consecutive symbols of $\mathcal{I}$ also occur in $I(y_p)$ to the right of $\mathcal{I}$ and to the left of the intervals of the phrases analyzed in Lemma 16, we claim that only $k-h$ such phrases can have $\phi(\cdot) > 1$. There are two cases: the h relevant symbols of y_p that do not overlap $\mathcal{I}$ are copied either from the beginning of $\mathcal{I}$ or from its end (the case $h = k$ trivially falls in both cases). If these symbols are copied from the beginning of $\mathcal{I}$, the first h phrases considered in the proof of Lemma 16 have $\phi(\cdot) \leq 1$, because their symbols can be divided into two substrings, the first of which can be copied from before $\mathcal{I}$, and the second from the portion of y_p following $\mathcal{I}$. If the symbols of y_p are copied from the last h symbols of $\mathcal{I}$, then the last h phrases considered in the proof of Lemma 16 have $\phi(\cdot) \leq 1$, because their symbols can be divided into two substrings, the first of which can be copied from preceding phrases of the same nature, and the second from the portion of y_p following $\mathcal{I}$.

Additionally, Lemma 11 produces a bound that is tight only for $k = 1$, and that can be further tightened.

Next, like in the proof of Theorem 3, we substitute a new expression for the terms $\frac{\overline{L}(\overline{L}+1)}{2} + \left\lceil \sqrt{\frac{(k+1)^2}{4} + (n-i-k+1)} - \frac{(k+1)}{2} \right\rceil$, the sum of a bound to maximum number of phrases pointing to the k changed symbols and a bound to the number of phrases whose RPOI contains the k changed symbols with padding at both ends that can have $\phi(\cdot) > 2$.

The new expression is constructed using the same pairing argument used in Theorem 3: we build the pseudo-parsing $\tilde{\mathbf{y}}$ of $\tilde{\mathbf{x}}$, an appropriately longer string than $\mathbf{x}$, pair each phrase of $\mathbf{y}$ having $\phi(\cdot) = 1$ with a phrase of $\tilde{\mathbf{y}}$, and each phrase of $\mathbf{y}$ not overlapping $\mathcal{I}$, with RPOI containing $\mathcal{I}$ with padding at both ends, and $\phi(\cdot) > 1$ with two phrases of $\tilde{\mathbf{y}}$. These two phrases of $\tilde{\mathbf{y}}$ have cumulative length equal to the length of the corresponding phrase of $\mathbf{y}$ plus k, as shown in Lemma 18. The arguments of Theorem 3 show that the number of phrases in $\tilde{\mathbf{y}}$ is larger than the number of phrases of $\mathbf{y}$ having $\phi(\cdot) = 1$ plus twice the number of phrases of $\mathbf{y}$ not overlapping $\mathcal{I}$, with RPOI containing $\mathcal{I}$ with padding at both ends, and $\phi(\cdot) > 1$. Lemma 13, and the remark on Definition 6 add $k-1$

to the upper bound. Lemma 16 adds $2k$. Lemma 14 further adds $k \log_2 k$. Finally, Lemma 11, accounts for the remaining $(k+3)$ additive term. □

7 Extensions

The techniques developed in the previous sections allow us to answer other questions related to the maximum expansion of a string compressed with LZ'77.

For example, we can bound the maximum expansion of a string of known compressibility that result from changing k consecutive symbols. As in the rest of the paper, we measure compressibility in terms of the number of phrases produced by the parsing and their length. Let the parsing $\mathbf{y}$ of a string $\mathbf{x}$ contain $\|\mathbf{y}\|$ phrases of length at least equal to 2, and assume that the symbols in the interval $[i, \ldots, i+k-1]$ are changed. Then a constructive lower bound on the expansion is $\min\left(\|\mathbf{y}\|, \underline{\Delta}_i^k\right)$. To obtain an upper bound, one would need to modify the lemmas of Section 5, and in particular the packing arguments, to account for the fact that the parsing of $\mathbf{x}$ contains exactly $\|\mathbf{y}\|$ of length equal to 2 or more.

We can also bound the maximum expansion of a string when k arbitrary symbols, non necessarily consecutive, are changed. In this case, the bounds can be derived by counting the maximum number of phrases that can point to any of the changed symbols, bounding the number of phrases with $\phi(\cdot) > 1$, and extending Lemma 14.

We can easily extend the results to the sliding-window LZ'77 of fixed length w, by recalling that a phrase can point only to previous occurrences that start within the window, and appropriately dealing with the boundary conditions, when compressing the first w symbols.

This paper contains bounds on the differences between the number of phrases of two parsings. Note that, although compressibility does not depend solely on the number of phrases in the parsing, we have chosen the number of phrases as the central quantity of our investigation, because we believe it provides a fundamental characterization that can be refined to individual specific schemes for mapping phrases into codewords. For example, in the work of Ziv and Lempel [1, Section II] assumes bounded delay decoding and yields a variable-to-fixed length encoding; our results can be easily adapted to this encoding approach, for example by constraining $\underline{L}$ and $\overline{L}$ of Theorem 4 and the length of phrases in Lemmas 16, 14, and 15 do not exceed the maximum encoding delay.

Finally, in Lemma 8 we assumed that the alphabet size was sufficiently large to ensure that

we can seed the string with $\underline{L} + \min(i, \underline{L}) + 1$ different symbols. An open problem is to relax the requirement on the alphabet size.

8 Discussion and Conclusions

We have introduced the problem of bounding the maximum expansion of a finite-length string parsed with LZ'77 that can result from changing $k \geq 1$ consecutive symbols.

We have proved upper and lower bounds to this maximum expansion as a function of the length n of the string, of the position i of the first changed symbol, and of the number k of changed symbols, under the assumption that the alphabet is sufficiently large. Although the assumption on the alphabet size limits the applicability of our (upper) bound in asymptotic analysis, it actually shows the relevance of the results to practical LZ'77 implementations, where the alphabet size is typically 256, and where the bounds in this paper are valid for sliding window sizes of up to 5.7 millions, well above those used in practice.

Further improvements to the upper bound for $k > 1$ may be possible via refinements of the $k \log_2 k$ term of Lemma 14, which is conjectured to be only linear in k.

The proofs of the lemmas suggest how the lower bound of Equation 8 and the upper bound of Equation 5 can be tightened to yield easily computable but less elegant solutions, described in Appendix A.

The results of this paper can be extended to cover the case in which $\mathbf{x}$ and $\hat{\mathbf{x}}$ differ in k non consecutive symbols and for the sliding-window LZ'77 algorithm with window size n, as well as for parallel, shared-dictionary LZ'77 algorithms: the necessary insights are provided by the lemmas proved herein.

These bounds are important to designing management policies for computer systems with main memory compression or compressed disk, where LZ'77-like algorithms are used to compress relatively small blocks of data, and where the compressibility can dramatically vary when a few symbols change.

Acknowledgments

This problem was proposed by Peter A. Franaszek, who independently constructed with Joy Thomas a lower bound similar to that of Theorem 2, and who provided us with valuable insight

on this topic during numerous enlightening discussions. We also wish to thank Professor Wojciech Szpankowski, for the discussions on how results such as those in the present paper could be used in conjunction with large deviation results for exploring theoretical bounds on redundancy.

A Numerical refinements of the bounds

We point out that the bounds obtained in this paper are somewhat loose. However, the proofs of the lemmas show immediately how to tighten the results to obtain easily computable tighter bounds, albeit not in closed form. The base of our proofs are the packing lemmas: we pack as many pseudo-phrases pointing to the symbols to be changed as we can. The results in this paper are based on crude upper and lower bounds to this number of pseudo-phrases. However, the actual number of pseudo-phrases can be easily obtained with a computer, and used to produce numerical upper and lower bounds that are in general substantially closer to each other than those shown in the theorem.

An example is given in Figure 1, which compares the upper and lower bounds derived in Theorem 2, the upper bound of Theorem 3, and numerically refined upper and lower bounds. The figure is for a string of 1024 symbols, and shows the bounds as a function of the position of the changed symbol.

The refined lower bound is obtained by noting that the right-hand side of Inequality (2) follows by constructing a string that parses into phrases of length 2 to $\underline{L}$ pointing to position i. If necessary, this string is padded with copies of $\tilde{\alpha}$ to ensure that it has length n. If the padding is of length ℓ_p, it can contain $\lfloor \ell_p/(\underline{L}+1) \rfloor \geq 0$ phrases of length $\underline{L}+1$ pointing to position i; the refined lower bound is computed by counting these phrases that are not accounted for in Theorem 2. Note that the ℓ_p can be equal to 0, in which case the refined bound coincides with the lower bound of Inequality (2).

The starting point for the refined upper bound is Theorem 3, which, for most values of i, tightens the upper bound of Theorem 2. Recall that in the proof of Theorem 3 we constructed a pseudo-string having length $m = n + \lfloor \sqrt{n-i+1} \rfloor$, packing to the right of position i in this pseudo-string two pseudo-phrases of length 2, three of length 3, etc., up to M pseudo-phrases of length M, where M is an approximate solution to Equation (6), and its value is precisely defined in the proof of the Theorem. There are three main sources of approximations in Theorem 3. The first is

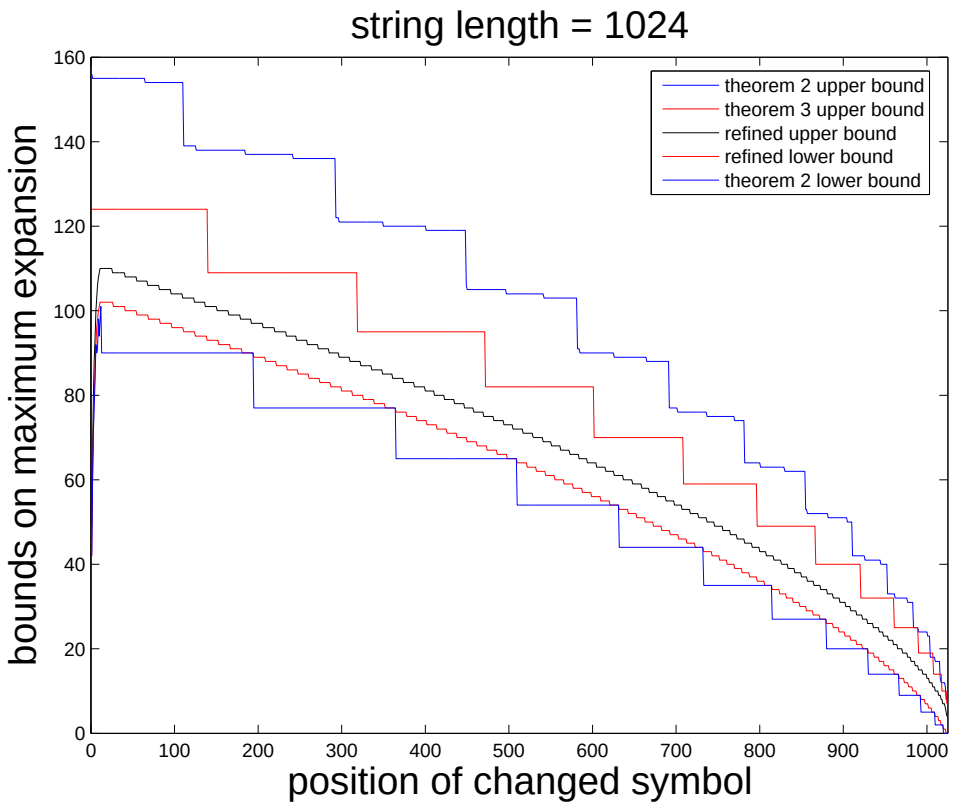


Figure 1: From top to bottom: the upper bound of Theorem 2, the upper bound of Theorem 3, the numerically refined upper bound, the numerically refined lower bound, and the lower bound of Theorem 2. The bounds are plotted as a function of the position of the changed symbol for a string of length 1024.

the value of M which is an approximation to the true solution of the cubic equation obtained using the technique described in Lemma 3. The second is that we count all pseudo-phrases of length up to M, even if the length of $\mathbf{x}$ is not sufficient to contain them all. The third is the edge effect for small values of i (accounted for in the lower bounds, e.g., in the proof of Lemma 8.: if i is small, there can be a phrase length ℓ^* such that not all possible RPOIs of phrases of length $\geq \ell^*$ can fit in the string $\mathbf{x}$.

Unlike in the case of the lower bound, where occasionally the refinement coincides with the original bound, for the refined upper bound of the figure the three refinements do not simultaneously coincide with the original results, except at $i = n = 1024$, and therefore the refined upper bound never coincides with the original one.

Further tightening of the bounds for the general $k \geq 1$ case can be obtained by refining Lemma 14, by noting that Lemmas 13, 16, 14, and 15 collectively over-count phrases that can have $\phi(\cdot) > 1$.

References

[1] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol. IT-23, no. 3, pp. 337–343, May 1977.

[2] A. Wyner and J. Ziv, "The sliding-window Lempel-Ziv algorithm is asymptotically optimal," *Proc. IEEE*, vol. 82, pp. 872–877, 1994.

[3] J. Ziv, "Coding theorems of individual sequences," *IEEE Trans. Inform. Theory*, vol. 24, no. 4, pp. 405–412, 1978.

[4] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 1, pp. 75–81, Jan. 1976.

[5] P. Shields, *The Ergodic Theory of Discrete Sample Paths*. Americal Mathematical Society, 1996.

[6] IBM Journal of Research and Development, "Special issue on memory compression," 45:2, 2001.

[7] P. Franaszek, J. Robinson, and J. Thomas, "Parallel compression with cooperative dictionary construction," in *Data Compression Conference, DCC'96*, 1996, pp. 200–209.

[8] B. Abali, M. Banikazemi, X. Shen, H. Franke, D. Poff, and B. Smith, "Hardware compressed main memory: Operating system support and performance evaluation," *IEEE Transactions on Computers*, vol. 50, no. 11, Nov. 2001.

[9] P. Franaszek and et. al., "Memory management method for preventing an operating system from writing into user memory space," US Patent 6,889,296.

[10] ——, "Reclaim space reserve for a compressed memory system," US Patent 6,842,832.

[11] ——, "Kernel identification for space management in compressed memory systems," US Patent 6,279,092.