Theoretical Computer Science 592 (2015) 115–134

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Approximation of grammar-based compression via recompression ☆

Artur Jeż

Institute of Computer Science, University of Wrocław, 50-383 Wrocław, Poland

ARTICLE INFO

Article history:
Received 7 November 2013
Received in revised form 15 May 2015
Accepted 22 May 2015
Available online 28 May 2015
Communicated by A. Apostolico

Keywords:
Grammar-based compression
Construction of the smallest grammar
SLP
Compression

ABSTRACT

In this paper we present a simple linear-time algorithm constructing a context-free grammar of size $\mathcal{O}(g \log(N/g))$ for the input string, where N is the size of the input string and g the size of the optimal grammar generating this string. The algorithm works for arbitrary size alphabets, but the running time is linear assuming that the alphabet Σ of the input string can be identified with numbers from $\{1, \ldots, N^c\}$ for some constant c. Otherwise, additional cost of $\mathcal{O}(N \log |\Sigma|)$ is needed.

Algorithms with such an approximation guarantee and running time are known, the novelty of this paper is a particular simplicity of the algorithm as well as the analysis of the algorithm, which uses a general technique of recompression recently introduced by the author. Furthermore, contrary to the previous results, this work does not use the LZ representation of the input string in the construction, nor in the analysis.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Grammar based compression

In the grammar-based compression text is represented by a context-free grammar (CFG) generating exactly one string. The idea behind this approach is that a CFG can compactly represent the structure of the text, even if this structure is not apparent. Furthermore, the natural hierarchical definition of the CFGs makes such a representation suitable for algorithms, in which case the string operations can be performed on the compressed representation, without the need of the explicit decompression [2,8,10,17,3,1]. Lastly, there is a close connection between block-based compression methods and the grammar compression: it is fairly easy to rewrite the LZW definition as an $\mathcal{O}(1)$ larger CFG, LZ77 can also be presented in this way, introducing a polynomial blow-up in size (reducing the blow up to $\log(N/\ell)$, where ℓ is the size of the LZ77 representation, is non-trivial [18,1]).

While grammar-based compression was introduced with practical purposes in mind and the paradigm was used in several implementations [12,11,16], it also turned out to be very useful in more theoretical considerations. Intuitively, in many cases large data have relatively simple inductive definition, which results in a grammar representation of a small size. On the other hand, it was already mentioned that the hierarchical structure of the CFGs allows operations directly on the compressed representation. A recent survey by Lohrey [13] gives a comprehensive description of several areas of theoretical computer science in which grammar-based compression was successfully applied.

☆ An extended abstract of this paper was presented at CPM 2013 conference [4].

E-mail address: aje@cs.uni.wroc.pl.

http://dx.doi.org/10.1016/j.tcs.2015.05.027
0304-3975/© 2015 Elsevier B.V. All rights reserved.

The main drawback of the grammar-based compression is that producing the smallest CFG for a text is *intractable*: given a string w and number k it is NP-hard to decide whether there exists a CFG of size k that generates w [20]. Furthermore, the size of the grammar cannot be approximated with an approximation factor better than $\frac{8569}{8568}$, unless P=NP [1].

Lastly, it is worth noting that in an extremely simple case of strings of the form $a^{\ell_1}ba^{\ell_2}b\cdots ba^{\ell_k}$ construction of the grammar generating a given string is equivalent (up to a small constant factor) to a construction of an *addition chain* for the sequence $\ell_1 < \ell_2 < \ldots < \ell_k$ and for the latter problem the best algorithm returns an addition chain of size $\log \ell_k + \mathcal{O}\left(\sum_{i=1}^{k} \frac{\log \ell_i}{\log\log \ell_i}\right)$ [21], which in particular yields an $\mathcal{O}\left(\frac{\log n}{\log\log n}\right)$ approximation of the size of the smallest addition chain. Since the addition chains are well studied, showing a construction of an addition chain shorter than $\log \ell_k + \mathcal{O}\left(\sum_{i=1}^{k} \frac{\log \ell_i}{\log\log \ell_i}\right)$ seems unlikely. Note, that this construction was not aimed at *approximating* the shortest addition chain, so it is still possible that $\mathcal{O}(\frac{\log n}{\log\log n})$ approximation can be improved. In any case, any new result for addition chains would be interesting on its own.

1.2. Approximation

The hardness of the smallest grammar problem naturally leads to two directions of research: on the one hand, several heuristics are considered [12,11,16], on the other, approximation algorithms, with a guaranteed approximation ratio, are proposed; in this paper we consider only the latter approach. Note that the heuristical algorithms can work differently depending on the distribution of letters in the input (and often the principle behind them assumes that the data has some sort of regularity). On the other hand, the approximation guarantees shown for the latter algorithms are universal, in the sense that they do not depend on the distribution of letters or any other properties of the provided text.

The first two algorithms with an approximation ratio $\mathcal{O}(\log(N/g))$ were developed independently (and simultaneously) by Rytter [18] and Charikar et al. [1]. They followed a similar approach, we first present Rytter's one as it is a bit easier to explain.

Rytter's algorithm [18] applies the LZ77 compression to the input string and then transforms the obtained LZ77 representation to an $\mathcal{O}(\ell \log(N/\ell))$ size grammar, where ℓ is the size of the LZ77 representation. It is easy to show that $\ell \leq g$ and as $f(x) = x\log(N/x)$ is increasing, the bound $\mathcal{O}(g\log(N/g))$ on the size of the grammar follows (and so a bound $\mathcal{O}(\log(N/g))$ on approximation ratio). The crucial part of the construction is the requirement that the intermediate constructed grammar defines a derivation tree satisfying the AVL condition. The bound on the running time and the approximation guarantee are all consequences of the balanced form of the derivation tree and of the known algorithms for merging, splitting, etc. of AVL trees (in fact these procedures are much simpler in this case, as we do not store any information in the internal nodes [18]). Note that also the final grammar for the input text is balanced, which makes it suitable for later processing. Since the construction of LZ77 representation can be performed in linear time (assuming that the letters of the input word can be sorted in linear time), also the running time of the whole algorithm can be easily bounded by a linear function.

Charikar et al. [1] followed more or less the same path, with a different condition imposed on the grammar: it was required that its derivation tree is length-balanced, i.e. for a rule $X \to YZ$ the lengths of words generated by Y and Z are within a certain multiplicative constant factor from each other. For such trees efficient implementation of merging, splitting etc. operations were given (i.e. constructed from scratch) by the authors and so the same running time as in the case of the AVL trees was obtained.

Lastly, Sakamoto [19] proposed a different algorithm, based on RePair [12], which is one of the practically implemented and used algorithms for grammar-based compression. His algorithm iteratively replaced pairs of different letters and maximal blocks of letters (a^ℓ is a *maximal block* if it cannot be extended by a to either side). A special pairing of the letters was devised, so that it is 'synchronising': if w has 2 disjoint occurrences in the text, then those two occurrences can be represented as $w_1w'w_2$, where $w_1, w_2 = \mathcal{O}(1)$, such that both occurrences of w' in text are paired and compressed in the same way. The analysis was based on considering the LZ77 representation of the text and proving that due to 'synchronisation' the factors of LZ77 are compressed very similarly as the text to which they refer.

However, to the author's best knowledge and understanding, the presented analysis [19] is incomplete, as the cost of nonterminals introduced when maximal blocks are replaced is not bounded at all in the paper, see the appendix; the bound that the author was able to obtain using the approach of Sakamoto is $\mathcal{O}(\log(N/g)^2)$, so worse than claimed.

1.3. Proposed approach: recompression

In this paper another algorithm is proposed, it is constructed using the general approach of *recompression*, developed by the author. In essence, we iteratively apply two replacement schemes to the text T:

pair compression of ab For two different symbols (i.e. letters or nonterminals) a, b such that substring ab occurs in T replace each of ab in T by a fresh nonterminal c.

a**'s block compression** For each maximal block a^ℓ, where a is a letter or a nonterminal and $\ell > 1$, that occurs in T, replace all a^ℓs in T by a fresh nonterminal a_ℓ.

Then the returned grammar is obtained by backtracking the compression operations performed by the algorithm: observe that replacing ab with c corresponds to a grammar production

$$c \to ab \tag{1a}$$

and similarly replacing a^ℓ with a_ℓ corresponds to a grammar production

$$a_\ell \to a^\ell \ . \tag{1b}$$

The algorithm is divided into *phases*: in the beginning of a phase, all pairs occurring in the current text are listed and stored in a list P, similarly, L contains all letters occurring in the current text. Then pair compression is applied to an appropriately chosen subset of P and all blocks of symbols from L are compressed, then the phase ends. If everything works perfectly, each symbol of T is replaced and so T's length drops by half; in reality the text length drops by some smaller, but constant, factor per phase. For the sake of simplicity, we treat all nonterminals introduced by the algorithm as letters.

In the previous work of the author it was shown that such an approach can be efficiently applied to text represented in a grammar compressed form. In this way new results for compressed membership problem [6], fully compressed pattern matching [8] and word equations [5,7] were obtained. In this paper a somehow opposite direction is followed: the recompression method is applied to the input string. This yields a simple linear-time algorithm: Performing one phase in $\mathcal{O}(|T|)$ running time is relatively easy, since the length of T drops by a constant factor in each phase, the $\mathcal{O}(N)$ running time is obtained. The used space is also linear, under the assumption that a unit machine word can hold $\Omega(\log n)$ bits.

However, the more interesting is the analysis, and not the algorithm itself: it is performed by applying (as a mental experiment) the recompression to the optimal grammar G for the input text. In this way, the current G always generates the current string kept by the algorithm and the number of nonterminals introduced during the construction can be calculated in terms of $|G| \le g$.

A relatively straightforward analysis yields that the generated grammar is of size $\mathcal{O}(g \log N)$, a slightly more involved analysis yields a bound $\mathcal{O}(g \log(N/g) + g)$ (for the same grammar).

1.4. Advantages and disadvantages of the proposed technique

The author believes that the proposed algorithm is interesting, as it is very simple and its analysis for the first time does not rely on LZ77 representation of the string. Potentially this can help in both design of an algorithm with a better approximation ratio and in showing a logarithmic lower bound: Observe that LZ77 representation is known to be at most as large as the smallest grammar, so it might be that some algorithm produces a grammar of size $o(g \log(N/g))$, even though this is of size $\Omega(\ell \log(N/\ell))$, where ℓ is the size of the LZ77 representation of the string. Secondly, as the analysis 'considers' the optimal grammar, it may be much easier to observe, where every approximation algorithm performs badly, and so try to approach a logarithmic lower bound. This is much harder to imagine, when the approximation analysis is done in terms of the LZ77.

Unfortunately, the obtained grammar is not balanced in any sense, in fact it is easy to give examples on which it returns grammar of height $\Omega(\sqrt{N})$ (note though that the same applies also to grammar returned by Sakamoto's algorithm). This makes the obtained grammar less suitable for later processing; on the other hand, the practically used grammar-based compressors [12,11,16] also do not produce a balanced grammar, nor do they give a guarantee on its height.

On the good side, there is no reason why the optimal grammar should be balanced, neither can we expect that for an unbalanced grammar a small balanced one exists. Thus it is possible that while $o(\log(N/g))$ approximation algorithm exists, there is no such an algorithm that always returns a balanced grammar.

We note that the reason why the grammar returned by the proposed algorithm can have large height is only due to block compression: if we assume that the nonterminal generating a^ℓ has height one, the whole grammar has height $\mathcal{O}(\log N)$. It looks reasonable to assume that many data structures for grammar representation of text as well as later processing of it can indeed process a production $a_\ell \to a^\ell$ in constant time.

Lastly, the proposed method seems to be much easier to generalise then the LZ77-based ones: generalisations of SLPs to grammars generating other objects (mostly: trees) are known but it seems that LZ77-based approach does not generalise to such settings, as LZ77 ignores any additional structure (like: tree-structure) of the data. In recent work of Lohrey and the author the algorithm presented in this paper is generalised to the case of tree-grammars, yielding a first provable approximation for the smallest tree grammar problem [9].

Comparison with Sakamoto's algorithm

The general approach is similar to Sakamoto's method, and both papers contain separate analyses and estimations for (variants of) pair compression and block compression. However, the pairing of letters seems more natural in the presented paper and the analysis is simpler. Also, the construction of nonterminals for blocks of letters is different, the author failed to show that the bound actually holds for the variant proposed by Sakamoto (see Appendix A). Note, that the analysis for block compression in this paper is much more involved than the one for pair compression. On the other hand, the connection to the addition chains suggests that the compression of blocks is the difficult part of the smallest grammar problem.

Note on computational model

The presented algorithm runs in linear time, assuming that the Σ can be identified with a continuous subset of natural numbers of size $\mathcal{O}(N^c)$ for some constant c and the RadixSort can be performed on it. Should this not be the case for the input, we can replace the original letters with such a subset, in $\mathcal{O}(n \log |\Sigma|)$ time (by creating a balanced tree for letters occurring in the input string). Note that the same comment applies to previous algorithms: there are many different algorithms for constructing the LZ77 representation of the text, but all of them first compute a suffix array (or a suffix tree) of the text, and linear-time algorithms for that are rely on linear-time sorting of letters (treated as integers); although Sakamoto's method was designed to work with constant-size alphabet, it can be easily extended to the case when Σ can be identified with a sequence of $\mathcal{O}(N^c)$ numbers, retaining the linear running-time.

In such a computational model we assume that a unit machine word can store a number using $\Theta(\log n)$ bits, and with such an assumption the space consumption of the algorithm is linear.

Roadmap

In Section 2 we define block and pair compression in detail and explain, how they are used to compute an SLP for a given string. We also describe, how to perform them in linear time: This is achieved by grouping appropriate substrings from a string, which is done using RadixSort. Then we show that the size of the input string is reduced by a constant factor in each phase, which can be used to conclude that the running time is linear.

Next we move to the analysis of the approximation ratio of the algorithm, which is performed in Section 3. To this end we consider the smallest grammar G generating this text, it is of size g. During the algorithm we modify appropriately the grammar, so that it always generates the current text T stored by the algorithm. The modifications follow the general recompression approach [6,8,5,7] and they are of the following type: when a nonterminal X of G generates awb we change its rule so that it now generates w and replace in each rule X by aXb, so that other nonterminals generate exactly the same strings. Moreover, only $\mathcal{O}(1)$ such modification per phase per variable are done, and so the size of the grammar increases by $\mathcal{O}(g)$ per phase. This is explained in Sections 3.4 and 3.5. The actual cost of the constructed rules is estimated based on the G stored by the analysis: whenever we perform a compression, we perform it also on the rules of G and thus the size of G decreases after each compression. The total cost is the size of all such decrements and it is the same as the total size of increments, which are $\mathcal{O}(g)$ per phase. There are only $\log N$ phases, which yields the approximation bound.

However, this analysis is unable to cover the cost introduced when very long blocks of letter a are replaced. We show that we can associate each such long block with a particular rule of a grammar G and so we charge the cost of such a representation to this rule. Then we show that if a rule is charged p (in total) then it originally produced a piece of input of length at least 2^p, thus at most $\log N$ is charged to a rule. This again yields $\mathcal{O}(g \log N)$ bound on the size of the generated rules.

Lastly, in Section 4 we improve the analysis to $\mathcal{O}(g + g\log(N/g))$. The crucial observation is that any reasonable grammar for a string T has size $\mathcal{O}(|T|)$, which follows by simple summation of lengths added by the productions. We divide the computation of the algorithm into two phases: the first one lasts until $|T| \approx g$ and the second begins afterwards and lasts till the end. We separately calculate the size of grammar productions introduced in those two phases. The latter are of size $\mathcal{O}(g)$, by the already mentioned observation. Using the same analysis as in the case of $\mathcal{O}(g \log N)$ we can show that the number of phases till the text is shortened to size g is $\mathcal{O}(\log(N/g))$ and so the size of the grammar is $\mathcal{O}(g\log(N/g))$. In most cases the generalisation is straightforward, with the only exception of the analysis for long blocks of a. In this case we need to make the analysis more precise: for each replaced power a^ℓ, instead of charging the cost to the rule, we mark ℓ letters of the input that somehow correspond to this block. We ensure that those markings are disjoint and their number of those markings is linear in g. Using standard calculus this yields a desired bound $\mathcal{O}(g\log(N/g))$. Details of this construction are separately presented in Section 4.7.

2. The algorithm

Notation conventions

The input sequence to be represented by a context-free grammar is $T \in \Sigma^*$, we shall use the same letter also for the text currently kept by the algorithm. By N we denote the initial length of T, by $|T|$ the current one. The algorithm TtoG introduces new symbols to the instance, those symbols are the nonterminals of the constructed grammar. However, these are later treated exactly as the original letters, so we insist on calling them *letters* as well and use common set Σ for both letters and nonterminals. We assume that T is represented as a doubly-linked list, so that removal and replacement of its elements can be performed in constant time (assuming that we have a link to such an occurrence). Note though that if we were to store T in a table, the running time would be the same. The smallest grammar generating the input sequence is denoted by G and its size $|G|$, measured as the length of the productions, is g.

Grammar

The crucial part of the analysis is the modification of G according to the compression performed on T. The terms non-terminal, rules, etc. always address the optimal grammar G (or its transformed version). To avoid confusion, we do not use terms 'production' and 'nonterminal' for a that replaced some substring in T (even though this is formally a nonterminal of the constructed grammar). Still, when a new 'letter' a is introduced to T we need to estimate the length of the 'productions' in the constructed grammar that are needed for a (note that we can of course use all letters previously used in T). The 'productions' introduced for a is called a *representation of a letter* a, the sum of lengths of those 'productions' is a *cost of representation of a letter* a (or simply: *representation cost*). For example, in production (1a) then the representation cost is 2 (as we have only one rule $c \to ab$) and in a rule (1b) we have a cost ℓ; the latter cost can be significantly reduced, for instance for a^{12} we can have a representation cost of 8 instead of 12, when we use a subgrammar $a_2 \to aa$, $a_3 \to a_2a$, $a_6 \to a_3a_3$ and $a_{12} \to a_6a_6$. Note that when c replaces a pair (as in (1a)), its representation cost is always 2, but when a replaces a block of letters, say a^ℓ, the cost might be larger than constant. In the latter case our algorithm constructs a special subgrammar for a_ℓ that generates a^ℓ, the exact way is given in Section 2.1.1.

2.1. The algorithm

The algorithm TtoG is divided into *phases*: in each phase we first list all letters and for each of them we perform the block compression and then again list all letters, choose appropriate partition and perform the pair compression for each pair from this partition that occurs in the text.

Algorithm 1 TtoG: outline.

while $|T| > 1$ **do**
  $L \leftarrow$ list of letters in T
  for each $a \in L$ **do** ▷ Blocks compression
    compress maximal blocks of a ▷ $\mathcal{O}(|T|)$
  $P \leftarrow$ list of pairs
  find partition of Σ into Σ_ℓ and Σ_r ▷ Covering at least 1/2 of occurrences of letters in T
  ▷ $\mathcal{O}(|T|)$, see Lemma 5
  for $ab \in P \cap \Sigma_\ell\Sigma_r$ **do** ▷ These pairs do not overlap
    compress pair ab ▷ Pair compression
return the constructed grammar

Before we make any analysis, we note that at the beginning of each phase we can make a linear-time preprocessing that guarantees that the letters in T form an interval of numbers (which makes them more suitable for sorting using RadixSort).

Lemma 1. *At the beginning of the phase, in time $\mathcal{O}(|T|)$ we can rename the letters used in T so that they form an interval of numbers.*

Proof. Observe that we assumed that the input alphabet consists of letters that can be identified with subset of $\{1, \dots, N^c\}$, see the discussion in the introduction. Treating them as vectors of length c over $\{0, \dots, N-1\}$ we can sort them using RadixSort in $\mathcal{O}(cN)$ time, i.e. linear one. Then we can re-number those letters to 1, 2, ..., n for some $n \leq N$.

Suppose that at the beginning of the phase the letters form an interval $[m, \dots, m+k]$. Each new letter, introduced in place of a compressed substring (i.e. a block a^ℓ or a pair ab), is assigned a consecutive number, and so after the phase the letters occurring in T are within an interval $[m..m+k']$ for some $k' > k$. It is now left to re-number the letters from $[m..m+k']$, so that the ones appearing in T indeed form an interval, which begins at $m+k'+1$. For each symbol a in the interval $[m..m+k']$ we set a flag to $\mathit{flag}[a] = 0$. Moreover, we set a variable *next* to $m+k'+1$. Then we read T. Whenever we spot a letter $a \in [m..m+k']$ with $\mathit{flag}[a] = 0$, we set $\mathit{flag}[a] := 1$; $\mathit{new}[a] := \mathit{next}$, and $\mathit{next} := \mathit{next} + 1$. Moreover, we replace this a by $\mathit{new}[a]$. When we spot a symbol $a \in [m..m+k']$ with $\mathit{flag}[a] = 1$, then we replace this a by $\mathit{new}[a]$. Clearly the running time is $\mathcal{O}(|T|)$ and after the algorithm the symbols form a subinterval $[m+k'+1..m+k'+k'']$ for appropriate $k'' \leq |T|$. □

2.1.1. Blocks compression

The blocks compression is very simple to implement: We read T, for a maximal block of as of length greater than 1 we create a record (a, ℓ, p), where ℓ is a length of the block, and p is the pointer to the first letter in this block. We then sort these records lexicographically using RadixSort (ignoring the last component). There are only $\mathcal{O}(|T|)$ records and we assume that Σ can be identified with an interval, see Lemma 1, this is all done in $\mathcal{O}(|T|)$. Now, for a fixed letter a, the consecutive tuples with the first coordinate a correspond to all blocks of a, ordered by the size. It is easy to replace them in $\mathcal{O}(|T|)$ time with new letters. Clearly, the space consumption is linear as well.

Lemma 2. *Block compression runs in $\mathcal{O}(\mathcal{O}(|T|))$ time and space.*

Note that so far we did not care about the cost of representation of new letters that replaced a-blocks. We use a particular schema to represent $a_{\ell_1}, a_{\ell_2}, \dots, a_{\ell_k}$, which shall have a representation cost $\mathcal{O}(\sum_{i=1}^{k}[1 + \log(\ell_i - \ell_{i-1})])$, where $\ell_0 = 0$ for convenience.

Lemma 3. *Given a list $1 < \ell_1 < \ell_2 < \cdots < \ell_k$ we can represent letters $a_{\ell_1}, a_{\ell_2}, \dots, a_{\ell_k}$ that replace blocks $a^{\ell_1}, a^{\ell_2}, \dots, a^{\ell_k}$ with a cost*

$$\mathcal{O}\left(\sum_{i=1}^{k}[1+\log(\ell_i-\ell_{i-1})]\right) ,$$

where $\ell_0 = 0$.

Proof. Firstly observe that without loss of generality we may assume that the list $\ell_1, \ell_2, \dots, \ell_k$ is given to us in a sorted way, as it can be easily obtained from the sorted list of occurrences of blocks. For simplicity define $\ell_0 = 0$ and let $\ell = \max_{i=1}^{k}(\ell_i - \ell_{i-1})$.

In the following, we shall define rules for certain new letters a_m, each of them 'derives' a^m (in other words, a_m represents a^m). For each $1 \le i \le \log \ell$ introduce a new letter a_{2^i}, defined as $a_{2^i} \to a_{2^{i-1}} a_{2^{i-1}}$, where a_1 simply denotes a. Clearly a_{2^i} represents a^{2^i} and the representation cost summed over all $i \le \ell$ is $2 \log \ell = \mathcal{O}(\log \ell)$.

Now introduce new letters $a_{\ell_i - \ell_{i-1}}$ for each $i > 0$, which shall represent $a^{\ell_i - \ell_{i-1}}$. They are represented using the binary expansion, i.e. by concatenation of at most $1 + \log(\ell_i - \ell_{i-1})$ from among the letters $a_1, a_2, a_4, \dots, a_{2^{\lfloor \log(\ell_i - \ell_{i-1}) \rfloor}}$. This has a representation cost $\mathcal{O}(\sum_{i=1}^{k}[1 + \log(\ell_i - \ell_{i-1})])$.

Lastly, each a_{ℓ_i} is represented as $a_{\ell_i} \to a_{\ell_i - \ell_{i-1}} a_{\ell_{i-1}}$, which has a total representation cost $\mathcal{O}(k)$.

Summing up $\mathcal{O}(\log \ell)$, $\mathcal{O}(\sum_{i=1}^{k}[1 + \log(\ell_i - \ell_{i-1})])$ and $\mathcal{O}(k)$ we obtain $\mathcal{O}(\sum_{i=1}^{k}[1 + \log(\ell_i - \ell_{i-1})])$. □

In the following we shall also use a simple property of the block compression: since no two maximal blocks of the same letter can be next to each other, after the block compression there are no blocks of length greater than 1 in T.

Lemma 4. *In* TtoG *right after the block compression, so in line 5, there are no two consecutive letters aa in T.*

Proof. Suppose for the sake of contradiction that there are such two letters. There are two cases:

- a **was present in** T **before block compression, so in line 2 of** TtoG But then a was listed in L in line 2 and aa was replaced by another letter in line 4.
- a **was introduced during block compression, so in line 4 of** TtoG Both a replaced some maximal blocks b^ℓ thus aa replaced $b^{2\ell}$, and so each of those two b^ℓs was not a maximal block. □

2.1.2. Pair compression

The pair compression is performed similarly as the block compression. However, since the pairs can overlap, compressing all pairs at the same time is not possible. Still, we can find a subset of non-overlapping pairs in T such that a constant fraction (1/4) of letters T is covered by occurrences of these pairs. This subset is defined by a *partition* of Σ into Σ_ℓ and Σ_r and choosing the pairs with the first letter in Σ_ℓ and the second in Σ_r; for a choice of $\Sigma_\ell \Sigma_r$ we say that occurrences of $ab \in P \cap \Sigma_\ell \Sigma_r$ are *covered* by $\Sigma_\ell \Sigma_r$.

Observe that the problem of finding such a partition reduces to the problem of finding a maximal weight cut in a directed weighted graph: for the reduction, we create a node for each letter and put an edge from a to b with weight k if there are k occurrences of pair ab in T. It is easy to see that a (directed) cut of weight k corresponds to a partition of letters covering exactly k occurrences of pairs and vice-versa. We use a standard solution to this problem [15, Section 6.3] presented in our terminology. We do that mostly for the reasons of running-time analysis.

The existence of a partition covering at least one fourth of the occurrences can be shown by a simple probabilistic algorithm: divide Σ into Σ_ℓ and Σ_r randomly, where each letter goes to each of the parts with probability 1/2. Consider two consecutive letters ab in T, note that they are different by Lemma 4. Then $a \in \Sigma_\ell$ and $b \in \Sigma_r$ with probability 1/4. There are $|T| - 1$ such pairs in T, so the expected number of pairs in T from $\Sigma_\ell \Sigma_r$ is $(|T| - 1)/4$. Observe, that if we were to count the number of pairs that are covered by $\Sigma_\ell \Sigma_r$ or by $\Sigma_r \Sigma_\ell$ then the expected number of pairs covered by $\Sigma_\ell \Sigma_r \cup \Sigma_r \Sigma_\ell$ is $(|T| - 1)/2$.

The deterministic construction of such a partition follows by a simple derandomisation [15, Section 6.3], using the conditional expectation approach. It is easier to first find a partition such that at least half of pairs' occurrences in T are covered by $\Sigma_\ell \Sigma_r \cup \Sigma_r \Sigma_\ell$ and then choose $\Sigma_\ell \Sigma_r$ or $\Sigma_r \Sigma_\ell$, depending on which of them covers more occurrences.

Lemma 5. *For T in $\mathcal{O}(|T|)$ time we can find (in line 6 of* TtoG*) a partition of Σ into Σ_ℓ, Σ_r such that number of occurrences of pairs $ab \in \Sigma_\ell \Sigma_r$ in T is at least $(|T| - 1)/4$.*

In the same running time we can provide, for each $ab \in P \cap \Sigma_\ell \Sigma_r$, a lists of pointers to occurrences of ab in T.

Proof. Suppose that we have already assigned some letters to Σ_ℓ and Σ_r and we are to decide, where the next letter a is assigned. If it is assigned to Σ_ℓ, then all occurrences of pairs from $a\Sigma_\ell \cup \Sigma_\ell a$ are not going to be covered, while occurrences of pairs from $a\Sigma_r \cup \Sigma_r a$ are; similarly observation holds for a being assigned to Σ_r. The algorithm makes a greedy choice, maximising the number of covered pairs in each step. As there are only two options, the choice brings in at least half of occurrences considered. Lastly, as each occurrence of a pair ab from T is considered exactly once (i.e. when the second of a, b is considered in the main loop), this procedure guarantees that at least half of occurrences of pairs in T is covered.

In order to make the selection effective, the algorithm GreedyPairs keeps an up to date counters $count_\ell[a]$ and $count_r[a]$, denoting, respectively, the number of occurrences of pairs from $a\Sigma_\ell \cup \Sigma_\ell a$ and $a\Sigma_r \cup \Sigma_r a$ in T (for the current assignment of letters to Σ_ℓ and Σ_r). Those counters are updated as soon as a letter is assigned to Σ_ℓ or Σ_r.

Algorithm 2 GreedyPairs.

1: $L \leftarrow$ set of letters used in P
2: $\Sigma_\ell \leftarrow \Sigma_r \leftarrow \emptyset$ ▷ Organised as a bit vector
3: **for** $a \in L$ **do**
4: $\quad count_\ell[a] \leftarrow count_r[a] \leftarrow 0$ ▷ Initialisation
5: **for** $a \in L$ **do**
6: $\quad$ **if** $count_r[a] \geq count_\ell[a]$ **then** ▷ Choose the one that guarantees larger cover
7: $\quad\quad choice \leftarrow \ell$
8: $\quad$ **else**
9: $\quad\quad choice \leftarrow r$
10: $\quad \Sigma_{choice} \leftarrow \Sigma_{choice} \cup \{a\}$
11: $\quad$ **for** each ab or ba occurrence in T **do**
12: $\quad\quad count_{choice}[b] \leftarrow count_{choice}[b] + 1$
13: **if** # occurrences of pairs from $\Sigma_r\Sigma_\ell$ in $T>$ # occurrences of pairs from $\Sigma_\ell\Sigma_r$ in T **then**
14: $\quad$ switch Σ_r and Σ_ℓ
15: **return** (Σ_ℓ, Σ_r)

By the argument given above, when Σ is partitioned into Σ_ℓ and Σ_r by GreedyPairs, at least half of the occurrences of pairs from T are covered by $\Sigma_\ell\Sigma_r \cup \Sigma_r\Sigma_\ell$. Then one of the choices $\Sigma_\ell\Sigma_r$ or $\Sigma_r\Sigma_\ell$ covers at least one fourth of the occurrences.

It is left to give an efficient variant of GreedyPairs, the non-obvious operations are the choice of the actual partition in lines 13–14 and the updating of $count_\ell[b]$ or $count_r[b]$ in line 12. All other operations clearly take at most $\mathcal{O}(|T|)$ time. The former is simple: since we organise Σ_ℓ and Σ_r as bit vectors, we can read T from left to right and calculate the number of pairs from $\Sigma_\ell\Sigma_r$ and those from $\Sigma_r\Sigma_\ell$ in $\mathcal{O}(|T|)$ time (when we read a pair ab we check in $\mathcal{O}(1)$ time whether $ab \in \Sigma_\ell\Sigma_r$ or $ab \in \Sigma_r\Sigma_\ell$). Afterwards we choose the partition that covers more occurrences of pairs in T.

To implement the *count*, for each letter a in T we have a *right list* $right(a) = \{b \mid ab \text{ occurs in } T\}$, represented as a list. Furthermore, the element b on right list $right(a)$ stores a list of all occurrences of the pair ab in T. There is a similar *left list* $left(a) = \{b \mid ba \text{ occurs in } T\}$. We comment, how to create left lists and right lists later.

Given *right* and *left*, performing the update in line 12 is easy: we go through $right(a)$ ($left(a)$) and increase the $count_\ell[b]$ (or $count_r[b]$) for each occurrence of ab (ba, respectively). Note that in this way each of the list $right(a)$ ($left(a)$) is read once during GreedyPairs, and so this time can be charged to their creation.

It remains to show how to initially create $right(a)$ ($left(a)$ is created similarly). We read T, when reading a pair ab we create a record (a, b, p), where p is a pointer to this occurrence. We then sort these records lexicographically using RadixSort. There are only $\mathcal{O}(|T|)$ records and we assume that Σ can be identified with an interval, see Lemma 1, so this all is done in $\mathcal{O}(|T|)$. Now, for a fixed letters a, the consecutive tuples with the first coordinate a can be turned into $right(a)$: for $b \in right(a)$ we want to store a list I of pointers to occurrences of ab, and on a sorted list of tuples the $\{(a, b, p)\}_{p\in I}$ are consecutive elements.

Lastly, in order to get for each $ab \in P \cap \Sigma_\ell\Sigma_r$ the lists of pointers to occurrences of ab in T it is enough to read *right* and filter the pairs such that $a \in \Sigma_\ell$ and $b \in \Sigma_r$; the filtering can be done in $\mathcal{O}(1)$ as Σ_ℓ and Σ_r are represented as bitvectors. The needed time is $\mathcal{O}(|T|)$.

The total running time is $\mathcal{O}(|T|)$, as each subprocedure has time constant per pair processed or $\mathcal{O}(|T|)$ in total. □

When for each pair $ab \in \Sigma_\ell\Sigma_r$ the list of its occurrences in T is provided, the replacement of pairs is done by going through the list and replacing each of the pair, which is done in linear time. Note, that as Σ_ℓ, Σ_r are disjoint, the considered pairs cannot overlap.

Lemma 6. *Pair compression runs in $\mathcal{O}(\mathcal{O}(|T|))$ time and space.*

2.2. Size and running time

It remains to estimate the total running time, summed over all phases. Clearly each subprocedure in a phase has a running time $\mathcal{O}(|T|)$ so it is enough to show that $|T|$ is reduced by a constant factor per phase.

Lemma 7. *In each phase $|T|$ is reduced by a constant factor.*

Proof. Let $m = |T|$ at the beginning of the phase. Let $m' \leq m$ be the length of T after the compression of blocks. First observe that if $m < 5$ then we satisfy the lemma when we make at least one compression, which can be always done, so in the following we assume that $m \geq 5$.

By Lemma 5 at least $(m' - 1)/4$ pairs are compressed during the pair compression, hence after this phase $|T'| \leq m' - (m' - 1)/4 \leq \frac{3}{4}m + \frac{1}{4}$. □

Theorem 1. TtoG *runs in linear time and linear space.*

Proof. Each phase clearly takes $\mathcal{O}(|T|)$ time and by Lemma 7 the $|T|$ drops by a constant factor in each phase. As the initial length of T is N, the total running time is $\mathcal{O}(N)$.

The space consumption is also linear. □

3. Size of the grammar: SLPs and recompression

To bound the cost of representation of letters introduced during the construction of the grammar, we start with the smallest grammar G generating (the input) T and then modify the grammar so that it generates T (i.e. the current string kept by TtoG) after each of the compression steps. Then the cost of representing the introduced letters is paid by various credits assigned to G. Hence, instead of the actual representation cost, which is difficult to estimate, we calculate the total value of issued credit. Note that this is entirely a mental experiment for the purpose of the analysis, as G is not stored or even known to the algorithm. We just perform some changes on it depending on the TtoG actions.

3.1. Definitions

We assume that grammar G is a *Straight Line Programme* (*SLP*), however, we relax the notion a bit (and call it an *SLP with explicit letters*, when an explicit reference is needed):

- the nonterminals are $X_1, \ldots, X_m$;
- each nonterminal has exactly one rule, which has at most two nonterminals in its body (i.e. there are two, one or none nonterminals and an arbitrary number of letters in the rule's body);
- if $X_i \to \alpha_i$ is a rule and X_j occurs in α_i then $j < i$.

Note that every CFG generating a unique string can be transformed into an SLP with explicit letters, with the size increased only by a constant factor:

- The renaming of nonterminals is obvious, we also remove the useless nonterminals.
- If a nonterminal X with a rule $X \to \alpha$ has more than two nonterminals in α, we can replace a substring YwZ in α by a new nonterminal X' with a rule $X' \to YwZ$. In this way the number of nonterminals in α drops by 1 and the size of the grammar increases by 1.
- As only one string is generated, we can reorder the nonterminals.

We call the letters (strings) occurring in the productions the *explicit letters* (*strings*, respectively). The unique string derived by X_i is denoted by $\text{val}(X_i)$; the grammar G shall satisfy the condition $\text{val}(X_m) = T$. We do not assume that $\text{val}(X_i) \neq \epsilon$, however, if $\text{val}(X_i) = \epsilon$ then X_i is not used in the productions of G (as this is a mental experiment, such X_i can be removed from the rules and in fact from the SLP).

3.2. Intuition and road map

3.2.1. Paying the representation cost: credit

With each explicit letter we associate two units of *credit* and pay most of the cost of representing the letters introduced during TtoG with these credits. More formally: when the algorithm modifies G and in the process it creates an occurrence of a letter, we *issue* (or pay) 2 new credits. On the other hand, if we do a compression step in G, then we remove some occurrences of letters. The credit associated with these occurrences is then *released* and can be used to pay for the representation cost of the new letters introduced by the compression step as well as for the credit for the newly introduced letters (so that the algorithm does not issue new credit). For pair compression the released credit indeed suffices to pay both the credit of the new letters occurrences and their representation cost, but for chain compression the released credit does not suffice, as it is not enough to pay the representation cost. Here we need some extra amount that is estimated separately later on in Section 3.6. In the end, the total cost is the sum of credit that was issued during the modifications of G plus the value that we estimate separately in Section 3.6.

3.2.2. Additional cost

The additional cost of representing letters during the block compression is estimated separately. In most cases, the cost of creating blocks can be cover by released credit, the only exception is when two long blocks of a are joined together. This can happen only between nonterminals in some rule of G and then the additional cost is charged towards this rule. Then we show that one rule has only $\mathcal{O}(\log N)$ cost charged to it: if we charge $\sum_i \log \ell_i$ cost to a rule, then it originally derived a word of length at least $\prod_i \ell_i$. This is described in detail in Section 3.6.

3.3. Modifying the grammar

Recall that whenever we say nonterminal, rule, production etc., we mean one of G.

When we replace each occurrence of the pair ab in T, we should also do this in G. However, this may not be possible, as some abs generated by G do not come from explicit pairs in G but rather are 'between' a nonterminal and a letter, for instance in a simple grammar $X_1 \to a$, $X_2 \to X_1 b$ the pair ab has such a problematic occurrence. If there are no such occurrences, it is enough to replace each explicit ab in G and we are done. To deal with the problematic ones, we need to somehow change the grammar, in the example above we replace X_1 with a, leaving only $X_2 \to ab$, for which the previous procedure can be applied. It turns out that this ad-hoc approach can be turned into a systematic procedure that deals with all such problems at once, by removing appropriate letters from rules and introducing them to other rules. This procedure is the main ingredient of this section and it is given in Section 3.4. Similar problems occur also when we want to replace maximal blocks of a and the solution to this problem is similar and it is given in Section 3.5.

Note that in the example above, when X_1 is replaced with a, 2 credit for the occurrence of a in $X_1 \to a$ is released and wasted. Then we issue 2 credit for the new occurrence of a in the rule X_2. When ab is replaced with c, 4 credit is released when ab is removed from the rule, 2 of this credit is used for the credit of c and the remaining 2 can be used to pay the representation cost for $c \to ab$.

3.4. Pair compression

A pair of letters ab has a *crossing occurrence* in a nonterminal X_i (with a rule $X_i \to \alpha_i$) if ab is in $\text{val}(X_i)$ but this occurrence does not come from an explicit occurrence of ab in α_i nor it is generated by any of the nonterminals in α_i. A pair is *non-crossing* if it has no crossing occurrence. Unless explicitly written, we use this notion only to pairs of *different* letters.

By $PC_{ab\to c}(w)$ we denote the text obtained from w by replacing each ab by a letter c (we assume that $a \neq b$). We say that a procedure (that changes a grammar G with nonterminals $X_1, \ldots, X_m$ to G' with nonterminals $X'_1, \ldots, X'_m$) *properly implements the pair compression* of ab to c, if $\text{val}(X'_m) = PC_{ab\to c}(\text{val}(X_m))$ and G' is an SLP with explicit letters. When a pair ab is noncrossing the procedure that implements the pair compression is easy to give: it is enough to replace each explicit ab with c.

Algorithm 3 PairCompNCr(ab, c): compressing a non-crossing pair ab.

1: replace each explicit ab in G by c

In order to distinguish between the nonterminals, grammar, etc. before and after the application of compression of ab (or, in general, any procedure) we use 'primed' letters, i.e. X'_i, G', T' for the nonterminals, grammar and text after this compression and 'unprimed', i.e. X_i, G, T for the ones before.

Lemma 8. *If ab is a noncrossing pair, then* PairCompNCr(ab, c) *properly implements the pair compression of ab. The credit of new letters in G' and cost of representing the new letter c is paid by the released credit; no new credit is issued. If a pair de, where $d \neq c \neq e$, is noncrossing in G, it is in G'.*

Proof. By induction on i we show that $\text{val}(X'_i) = PC_{ab\to c}(\text{val}(X_i))$. Consider any occurrence of ab in the string generated by X_i. If it is an explicit string then it is replaced by PairCompNCr(ab, c). If it is contained within substring generated by some X_j, this occurrence was compressed by the inductive assumption. The remaining case is the crossing occurrence of ab: since the only modifications to the rules made by PairCompNCr(ab, c) is the replacement of ab by c, such a crossing pair existed already before PairCompNCr(ab, c), but this is not possible by the lemma assumption that ab is non-crossing.

Each occurrence of ab had 4 units of credit while c has only 2, so the replacement released 4 units of credit, 2 of which are used to pay for the credit of c and the other 2 to pay the cost of representation of c (if we replace more than one occurrence of ab then some credit is wasted).

Lastly, replacing ab in G by a new letter c cannot make de (where $d \neq c \neq e$) a crossing pair in G, as no new occurrence of d, e was introduced on the way. □

If all pairs in $\Sigma_\ell \Sigma_r$ are non-crossing, iteration of PairCompNCr(ab, c) for each pair ab in $\Sigma_\ell \Sigma_r$ properly implements the pair compression for all pairs in $\Sigma_\ell \Sigma_r$ (note that as Σ_ℓ and Σ_r are disjoint, occurrences of different pairs from $\Sigma_\ell \Sigma_r$ cannot overlap and so the order of replacement does not matter). So it is left to assure that indeed the pairs from $\Sigma_\ell \Sigma_r$ are all noncrossing. It is easy to see that $ab \in \Sigma_\ell \Sigma_r$ is a crossing pair if and only if one of the following three 'bad' situations occurs:

CP1 there is a nonterminal X_i, where $i < m$, such that $\text{val}(X_i)$ begins with b and aX_i occurs in one of the rules;
CP2 there is a nonterminal X_i, where $i < m$, such that $\text{val}(X_i)$ ends with a and $X_i b$ occurs in one of the rules;
CP3 there are nonterminals X_i, X_j, where $i, j < m$, such that $\text{val}(X_i)$ ends with a and $\text{val}(X_j)$ begins with b and $X_i X_j$ occurs in one of the rules.

Consider (CP1), let $bw = \text{val}(X_i)$. Then it is enough to modify the rule for X_i so that $\text{val}(X_i) = w$ and replace each X_i in the rules by bX_i, we call this action the *left-popping b from X_i*. Similar operation of right-popping a letter a from X_i is symmetrically defined. It is shown in Lemma 9 below that they indeed take care of all crossing occurrences of ab.

Furthermore, left-popping and right-popping can be performed for many letters in parallel: the below procedure Pop(Σ_ℓ, Σ_r) 'uncrosses' all pairs from the set $\Sigma_\ell \Sigma_r$, assuming that Σ_ℓ and Σ_r are disjoint subsets of Σ (and we apply Pop(Σ_ℓ, Σ_r) only in the cases in which they are).

Algorithm 4 Pop(Σ_ℓ, Σ_r): Popping letters from Σ_ℓ and Σ_r.

```
1:  for i ← 1..m − 1 do
2:      let the production for X_i be X_i → α_i
3:      if the first symbol of α_i is b ∈ Σ_r then          ▷ Left-popping b
4:          remove this b from α_i
5:          replace X_i in G's productions by bX_i
6:          if val(X_i) = ϵ then
7:              remove X_i from G's productions
8:  for i ← 1..m − 1 do
9:      let the production of X_i be X_i → α_i
10:     if the last symbol of α_i is a ∈ Σ_ℓ then           ▷ Right-popping a
11:         remove this a from α_i
12:         replace X_i in G's productions by X_i a
13:         if val(X_i) = ϵ then
14:             remove X_i from G's productions
```

Lemma 9. *After application of* Pop(Σ_ℓ, Σ_r)*, where $\Sigma_\ell \cap \Sigma_r = \emptyset$, none of the pairs $ab \in \Sigma_\ell \Sigma_r$ is crossing. Furthermore,* $\text{val}(X'_m) = \text{val}(X_m)$*. At most $\mathcal{O}(m)$ credit is issued during* Pop(Σ_ℓ, Σ_r).

Proof. Observe first that whenever we remove b from the front of some α_i we replace each of X_i occurrence with bX_i and if afterwards $\text{val}(X_i) = \epsilon$ then we remove X_i from the rules, hence the words derived by each other nonterminal (in particular X_m) do not change, the same applies to replacement of X_i with $X_i a$. Hence, in the end $\text{val}(X'_m) = \text{val}(X_m) = T$ (note that we do not pop letters from X_m).

Secondly, we show that if $\text{val}(X'_i)$ begins with a letter $b' \in \Sigma_r$ then we left-popped a letter from X_i (which by the code is some $b \in \Sigma_r$), a similar claim (by symmetry) of course holds for the last letter of $\text{val}(X_i)$ and Σ_ℓ. So suppose that the claim is not true and consider the nonterminal X_i with the smallest i such that $\text{val}(X'_i)$ begins with $b' \in \Sigma_r$ but we did not left-pop a letter from X_i. Consider what was the first symbol in α_i when Pop considered X_i in line 3. As Pop did not left-pop a letter from X_i, the first letter of $\text{val}(X_i)$ and $\text{val}(X'_i)$ is the same and hence it is $b' \in \Sigma_r$. So α_i cannot begin with a letter as then it is $b' \in \Sigma_r$, which should have been left-popped. Hence it is some nonterminal X_j for $j < i$. But then $\text{val}(X'_j)$ begins with $b' \in \Sigma_r$ and so by the induction assumption Pop left-popped a letter from X_j. But there was no way to remove this letter from α_i, so α_i should begin with a letter, contradiction.

Suppose that after Pop there is a crossing pair $ab \in \Sigma_\ell \Sigma_r$. There are three already mentioned cases (CP1)–(CP3): consider only (CP1), in which aX_i occurs in the rule and $\text{val}(X_i)$ begins with b. Note that as $a \notin \Sigma_r$ is the letter to the left of X'_i, X'_i did not left-pop a letter. But it begins with $b \in \Sigma_r$, so it should have. Contradiction. The other cases are dealt with in a similar manner.

Note that at most 4 new letters are introduced to each rule (2 per nonterminal) thus at most $8m$ credit is issued. □

In order to compress pairs from $\Sigma_\ell \Sigma_r$ it is enough to first uncross them all using Pop(Σ_ℓ, Σ_r) and then compress them all by PairCompNCr(ab, c) for each $ab \in \Sigma_\ell \Sigma_r$.

Algorithm 5 PairComp(Σ_ℓ, Σ_r): compresses pairs from $\Sigma_\ell \Sigma_r$.

```
run Pop(Σℓ, Σr)
for ab ∈ ΣℓΣr do
    run PairCompNCr(ab, c)                              ▷ c is a fresh letter
```

Lemma 10. PairComp *implements pair compression for each $ab \in \Sigma_\ell \Sigma_r$. It issues $\mathcal{O}(m)$ new credit to G, where m is the number of nonterminals of G. The credit of the new letters introduced to G and their representation costs are covered by the credit issued or released by* PairComp.

Proof. By Lemma 9 after Pop(Σ_ℓ, Σ_r) each pair in $\Sigma_\ell \Sigma_r$ is non-crossing and $\mathcal{O}(m)$ credit is issued in the process, furthermore val(X_m) does not change.

By Lemma 8 for a non-crossing pair ab the PairCompNCr(ab, c) implements the pair compression, furthermore, any other non-crossing pair $a'b' \in \Sigma_\ell \Sigma_r$ remains non-crossing. Lastly, all occurrences of different pairs from $\Sigma_\ell \Sigma_r$ are disjoint (as Σ_ℓ and Σ_r are disjoint subsets of Σ) as so the order of replacing them does not matter and so we implemented the pair compression for all pairs in $\Sigma_\ell \Sigma_r$. The cost of representation and credit of new letters is covered by the released credit, see Lemma 8. □

Using Lemma 10 we can estimate the total credit issued during the pair compression.

Corollary 1. *The compression of pairs issues in total $\mathcal{O}(m \log N)$ credit during the run of* TtoG*; the credit of the new letters introduced to G and their representation costs are covered by the credit issued or released during* PairComp.

3.5. Blocks compression

Similar notions and analysis as the ones for pairs are applied for blocks. Consider occurrences of maximal a-blocks in T and their derivation by G. Then a block a^ℓ has a *crossing occurrence* in X_i with a rule $X_i \to \alpha_i$, if it is contained in val(X_i) but this occurrence is not generated by the explicit as in the rule nor in the substrings generated by the nonterminals in α_i. If a-blocks have no crossing occurrences, then a *has no crossing blocks*. As for noncrossing pairs, the compression of a blocks, when it has no crossing blocks, is easy: it is enough to replace each explicit maximal a-block in the rules of G. We use similar terminology as in the case of pairs: we say that a subprocedure properly implements a block compression for a.

Algorithm 6 BlockCompNCr(a), which compresses a blocks when a has no crossing blocks.

```
for each a^{ℓm} do
    replace every explicit maximal block a^{ℓm} in G by a_{ℓm}
```

Lemma 11. *If a has no crossing blocks then* BlockCompNCr(a) *properly implements the a's blocks compression.*

Furthermore, if a letter b from T had no crossing blocks in G, it does not have them in G'.

The proof is similar to the proof of Lemma 8 and so it is omitted. Note that we do not yet discuss the issued credit, nor the cost of the representation of letters representing blocks (the latter is done in Section 3.6).

It is left to ensure that no letter has a crossing block. The solution is similar to Pop, this time though we need to remove the whole prefix and suffix from val(X_i) instead of a single letter. The idea is as follows: suppose that a has a crossing block because aX_i occurs in the rule and val(X_i) begins with a. Left-popping a does not solve the problem, as it might be that val(X_i) still begins with a. Thus, we keep on left-popping until the first letter of val(X_i) is not a, i.e. we remove the a-prefix of val(X_i). The same works for suffixes.

Algorithm 7 RemCrBlocks: removing crossing blocks.

```
for i ← 1..m − 1 do
    let a, b be the first and last letter of val(X_i)
    let ℓ_i, r_i be the length of the a-prefix and b-suffix of val(X_i)
                                  ▷ If val(X_i) ∈ a* then r_i = 0 and ℓ_i = |val(X_i)|
    remove a^{ℓ_i} from the beginning and b^{r_i} from the end of α_i
    replace X_i by a^{ℓ_i} X_i b^{r_i} in the rules
    if val(X_i) = ϵ then
        remove X_i from the rules
```

Lemma 12. *After* RemCrBlocks *no letter has a crossing block and* val(X_m) = val(X'_m).

Proof. Firstly, $\text{val}(X'_m) = \text{val}(X_m)$: observe that when we remove a-prefix a^{ℓ_i} from α_i we replace each X_i with $a^{\ell_i} X_i$ (and similarly for the b-suffix), also when we remove X_i from the rules then $\text{val}(X_i) = \epsilon$. Hence when processing X_i, the strings generated by all other nonterminals are not affected. In particular, as we do not remove the prefix and suffix of X_m, the string generated by X_m remains the same after RemCrBlocks.

By above observation, the value of $\text{val}(X_i)$ does not change until RemCrBlocks considers X_i. We show that when RemCrBlocks considers X_i such that $\text{val}(X_i)$ has a-prefix a^{ℓ_i} and b-suffix b^{r_i}, then α_i begins with a^{ℓ_i} and ends with b^{r_i} (the trivial case, when $\text{val}(X_i) = a^{\ell_i}$ is shown in the same way). Suppose that this is not the case and consider X_i with smallest i for which this is not true. Clearly it is not X_1, as there are no nonterminals in α_1 and so $\text{val}(X_1) = \alpha_1$. So let X_i have a rule $X_i \to \alpha_i$, we deal only with the a-prefix, the proof of b-suffix is symmetrical. Since the a-prefix of $\text{val}(X_i)$ and α_i are different, this means that the a-prefix of $\text{val}(X_i)$ is partially generated by the first nonterminal in α_i, let it be X_j. By the choice of i we know that X_j popped its prefix (of some letter, say a') and so it was replaced with $a'^{\ell_j} X'_j$. Furthermore, $\text{val}(X'_j)$ begins with $a'' \neq a'$. Since there is no way to remove this a' prefix from α_i, this a'^{ℓ_j} is part of the a-prefix of $\text{val}(X_i)$, in particular $a' = a$. However, $\text{val}(X'_j)$ begins with $a'' \neq a$, so the a-prefixes of α_i and $\text{val}(X_i)$ are the same, contradiction.

As a consequence, if aX_i occurs in any rule after RemCrBlocks, then a is not the first letter of $\text{val}(X_i)$, as prefix of letters a was removed from X_i. Other cases are handled similarly. So there are no crossing blocks after RemCrBlocks. □

The compression of all blocks of letters is done by first running RemCrBlocks and then compressing each of the block by BlockCompNCr. Note that we do not compress blocks of letters that are introduced in this way. Concerning the number of credit, the arbitrary long blocks popped by RemCrBlocks are compressed (each into a single letter) and so at most 8 credit per rule is issued.

Algorithm 8 BlockComp: compresses blocks of letters.

```
1: run RemCrBlocks
2: L ← list of letters in T
3: for each a ∈ L do
4:     run BlockCompNCr(a)
```

Lemma 13. BlockComp *properly implements the blocks compression for each letter a occurring in T before its application and issues $\mathcal{O}(m)$ credit. The issued credit covers the cost of credit of letters introduced during the* BlockComp *(but not their representation cost).*

The proof is similar as the proof of Lemma 10 so it is omitted.

Corollary 2. *During the whole* TtoG *the* BlockComp *issues in total $\mathcal{O}(m \log N)$ credit. The credit of the new letters introduced to G is covered by the issued credit.*

Note that the cost of representation of letters replacing blocks is not covered by the credit, this cost is separately estimated in the next subsection.

3.6. Calculating the cost of representing letters in block compression

The issued credit is enough to pay the 2 credit for occurrences of letters introduced during TtoG and the released credit is enough to pay the credit of the letters introduced during the pair compression and their representation cost. However, credit alone cannot cover the representation cost of letters replacing blocks. The appropriate analysis is presented in this section. The overall plan is as follows: firstly, we define a scheme of representing the letters based on the grammar G and the way G is changed by BlockComp (the *G-based representation*). Then for such a representation schema, we show that the cost of representation is $\mathcal{O}(g \log N)$. Lastly, it is proved that the actual cost of representing the letters by TtoG (the TtoG-*based representation*) is smaller than the G-based one, hence it is also $\mathcal{O}(g \log N)$.

3.6.1. G-based representation

The intuition is as follows: while the a blocks can have exponential length, most of them do not differ much, as in most cases the new blocks are obtained by concatenating letters a that occur explicitly in the grammar and in such a case the released credit can be used to pay for the representation cost. This does not apply when the new block is obtained by concatenating two different blocks of a (popped from nonterminals) inside a rule. However, this cannot happen too often: when blocks of length $p_1, p_2, \ldots, p_\ell$ are compressed (at the cost of $\mathcal{O}\Big(\sum_{i=1}^{\ell} (1 + \log p_i)\Big) = \mathcal{O}(\log(\prod_{i=1}^{\ell} p_i))$, as each $p_i \geq 2$), the length of the corresponding text in the input text is $\prod_{i=1}^{\ell} p_i$, which is at most N. Thus $\mathcal{O}\Big(\sum_{i=1}^{\ell} (1 + \log p_i)\Big) = \mathcal{O}(\log \prod_{i=1}^{\ell} p_i) = \mathcal{O}(\log N)$ cost per nonterminal is scored.

Getting back to the representation of letters: we create a new letter for each a block in the rule $X_i \to \alpha_i$ after RemCrBlocks popped prefixes and suffixes from $X_1, \ldots, X_{i-1}$ but before it popped letters from X_i. (We add the artificial empty block ϵ to streamline the later description and analysis.) Such a block is a *power* if it is obtained by concatenation of two a-blocks popped from nonterminals inside a rule (and perhaps some other explicit letters a), note that this power may be then popped from a rule (as it may be a prefix or suffix in this rule). This implies that in the rule $X_i \to uX_jvX_kw$ the popped suffix of X_j and popped prefix of X_k are blocks of the same letter, say a, and furthermore $v \in a^*$. Note that it might be that one (or both) of X_j and X_k were removed in the process (in this case the power can be popped from a rule as well). For each block a^ℓ that is not a power we may uniquely identify another block a^k (perhaps ϵ, not necessarily a power) such that a^ℓ was obtained by concatenating $\ell - k$ explicit letters to a^k in some rule.

Lemma 14. *For each block a^ℓ represented in the G-based representation that is not a power there is block a^k (perhaps $k = 0$) such that a^k is also represented in G-based representation and a^ℓ was obtained in a rule by concatenating $\ell - k$ explicit letters that existed in the rule to a^k.*

Note that the block a^k is not necessarily unique: it might be that there are several a^ℓ blocks in G which are obtained as different concatenations of a^k and $\ell - k$ explicit letters.

Proof. Let a_ℓ be created in the rule for X_i, after popping prefixes and suffixes from $X_1, \dots, X_{i-1}$. Consider, how many popped prefixes and suffixes take part in this a^ℓ.

If two, then it is a power, contradiction.

If one, then let the popped prefix (or suffix) be a^k. Since it was popped, say from X_j, then a^k was a maximal block in X_j before popping, so it is represented as well. Then in the rule for X_i the a^ℓ is obtained by concatenating $\ell - k$ letters a to a^k. None of those letters come from popped prefixes and suffixes, so they are all explicit letters that were present in this rule.

If there are none popped prefixes and suffixes that are part of this a^ℓ, then all its letters are explicit letters from the rule for X_i, and we treat it as a concatenation of k explicit letters to ϵ. □

We represent the blocks as follows:

1. for a block a^ℓ that is a power we represent a_ℓ using the binary expansion, which costs $\mathcal{O}(1 + \log \ell)$;
2. for a block a^ℓ that is obtained by concatenating $\ell - k$ explicit letters to a block a^k (see Lemma 14) we represent a_ℓ as $a_k \underbrace{a \cdots a}_{\ell - k \text{ times}}$, which has a representation cost of $\ell - k + 1$, this cost is covered by the $2(\ell - k) \geq \ell - k + 1$ credit released by the $\ell - k$ explicit letters a. Note that the credit released by those letters was not used for any other purpose. (Furthermore recall that the 2 units of credit per occurrence of a_ℓ in the rules of grammar are already covered by the credit issued by BlockComp, see Lemma 13.)

We refer to cost in 1 as the *cost of representing powers* and redirect this cost to the nonterminal in whose rule this power is created. The cost in 2, as marked there, is covered by released credit.

3.6.2. Cost of G-based representation

We now estimate the cost of representing powers. The idea is that if nonterminal X_i is charged the cost of representing powers of length $p_1, p_2, \dots, p_\ell$, which have representation cost $\mathcal{O}(\sum_{i=1}^{\ell} 1 + \log p_i) = \mathcal{O}(\log(\prod_{i=1}^{\ell} p_i))$, then in the input this nonterminal generated a text of length at least $p_1 \cdot p_2 \cdots p_\ell \leq N$ and so the total cost of representing powers is $\mathcal{O}(\log N)$ (per nonterminal). This is formalised in the lemma below.

Lemma 15. *The total cost of representing powers by G-based representation charged towards a single rule is $\mathcal{O}(\log N)$.*

Proof. There are two cases: first, after the creation of the power in a rule $X_i \to uX_jvX_kw$ one of the nonterminals X_j, X_k is removed. But this happens at most once for the rule and the cost of $\mathcal{O}(\log N)$ of representing the power can be charged to a rule.

The second and crucial case is when after the creation of power both nonterminals remained in a rule $X_i \to uX_jvX_kw$. Note that creation of the a power here means that $\text{val}(X_j)$ has a-suffix, $\text{val}(X_k)$ an a-prefix and $v \in a^*$.

Fix this rule and consider all such creations of powers performed on this rule. Let the consecutive letters, whose blocks are compressed, be $a^{(1)}, a^{(2)}, \dots, a^{(\ell)}$ and their lengths $p_1, p_2, \dots, p_\ell$. Lastly, the p_ℓ repetitions of $a^{(\ell)}$ are replaced by $a^{(\ell+1)}$. (Observe, that $a^{(i+1)}$ does not need to be the letter that replaced the $a^{(i)}$'s block, as there might have been some other compression performed on that letter.) Then the cost of the representing powers is constant time more than

$$\sum_{i=1}^{\ell} (1 + \log p_i) \leq 2 \sum_{i=1}^{\ell} \log p_i \quad . \tag{2}$$

Define *weight*: for a letter it is the length of the substring of the original input string that it 'derives'. Note that the maximal weight of any letter is N, the length of the input word.

Consider the weight of the strings between X_j and X_k. Clearly, after the i-th blocks compression it is exactly $p_i \cdot \mathrm{w}(a^{(i)})$, as the block of p_i letters $a^{(i)}$ was replaced by one letter. We claim that $\mathrm{w}(a^{(i+1)}) \geq p_i \, \mathrm{w}(a^{(i)})$: right after the i-th blocks compression the string between X_j and X_k is simply a letter $a^{(i)}_{p_i}$, which replaced the p_i block of $a^{(i)}$. After some operations, this string consists of p_{i+1} letters $a^{(i+1)}$. Observe that $(a^{(i+1)})^{p_{i+1}}$ 'derives' $a^{(i)}_{p_i}$: indeed all operations performed by TtoG do not remove the letters from string between X_j and X_k in a rule, only replace strings with single letters and perhaps add letters at the ends of this string. But if $(a^{(i+1)})^{p_{i+1}}$ 'derives' $a^{(i)}_{p_i}$, i.e. a single letter, then also $a^{(i+1)}$ 'derives' $a^{(i)}_{p_i}$, hence

$$\mathrm{w}(a^{(i+1)}) \geq \mathrm{w}(a^{(i)}_{p_i}) = p_i \, \mathrm{w}(a^{(i)}) \ .$$

Since $\mathrm{w}(a^{(1)}) \geq 1$ it follows that $\mathrm{w}(a^{(\ell+1)}) \geq \prod_{i=1}^{\ell} p_i$. As $\mathrm{w}(a^{(\ell+1)}) \leq N$ we have

$$N \geq \prod_{i=1}^{\ell} p_i$$

and so it can be concluded that

$$\begin{aligned} \log(N) &\geq \log\left(\prod_{i=1}^{\ell} p_i\right) \\ &= \sum_{i=1}^{\ell} \log p_i \ . \end{aligned}$$

Therefore, the whole cost $\sum_{i=1}^{\ell} \log p_i$, as estimated in (2), is $\mathcal{O}(\log N)$, as claimed. □

Corollary 3. *The cost of G-based representation is $\mathcal{O}(g + g \log N)$.*

Proof. Concerning the cost of representing powers, by Lemma 15 we redirect at most $\mathcal{O}(\log N)$ against each of the $m \leq g$ rules of G. The cost of representing non-powers is covered by the released credit; the initial value of credit is at most $2g$ and by Corollary 1 and Corollary 2 at most $\mathcal{O}(g \log N)$ credit is issued during the whole run of TtoG, which ends the proof. □

3.6.3. Comparing the G-based representation cost and TtoG*-based representation cost*

We now show that the cost of TtoG-based representation is at most as high as G-based one. We first represent G-based representation cost using a weighted graph $\mathcal{G}_G$, such that the G-based representation is (up to a constant factor) $w(\mathcal{G}_G)$, i.e. the sum of weights of edges of $\mathcal{G}_G$.

Lemma 16. *The cost of G-based representation of all blocks is $\Theta(w(\mathcal{G}_G))$, where nodes of $\mathcal{G}_G$ are labelled with blocks represented in the G-based representation and edge from a^ℓ to a^k, where $\ell > k$, has weight $\ell - k$ or $1 + \log(\ell - k)$ (in this case additionally $k = 0$). Each node (other than a and ϵ) has at least one outgoing edge.*

The former corresponds to the representation cost covered by the released credit while the latter to the cost of representing powers.

Proof. We give a construction of the graph $\mathcal{G}_G$.

Fix the letter a and consider any of the blocks a^ℓ that is represented by G, we put a node a^ℓ in $\mathcal{G}_G$. Note that a single a^ℓ may be represented in many ways: different occurrences of a^ℓ are replaced with a_ℓ and may be represented in different ways (or even twice in the same way), this means that $\mathcal{G}_G$ may have more than one outgoing edge per node.

- when a^ℓ is a power, we create an edge from the node labelled with a^ℓ to ϵ, the weight is $1 + \log \ell$ (recall that this is the cost of representing this power);
- when a_ℓ is represented as a concatenation of $\ell - k$ letters to a_k, we create and edge from the node a^ℓ to a^k, the weight is $\ell - k$ (this is the cost of representing this block; it was paid by the credit on the $\ell - k$ explicit letters a).

Then the sum of the weight of the created graph is a cost of representing the blocks using the G-based representation (up to a constant factor). □

Similarly, the cost of TtoG-based representation has a graph representation $\mathcal{G}_{\mathsf{TtoG}}$.

Lemma 17. *The cost of* TtoG*-representation for blocks of a letter a is $\Theta(w(\mathcal{G}_{\mathsf{TtoG}}))$, where the nodes of $\mathcal{G}_{\mathsf{TtoG}}$ are labelled with blocks represented by* TtoG*-representation and it has an edge from a^ℓ to a^k if and only if ℓ and k are two consecutive lengths of a-blocks. Such an edge has weight $1 + \log(\ell - k)$.*

Proof. Observe that this is a straightforward consequence of the way the blocks are represented: Lemma 3 guarantees that when blocks $a^{\ell_1}, a^{\ell_2}, \ldots, a^{\ell_k}$ (where $1 < \ell_1 < \ell_2 < \cdots < \ell_k$) are represented the TtoG-representation cost is $\mathcal{O}(\sum_{i=1}^{k}[1 + \log(\ell_i - \ell_{i-1})])$, so we can assign cost $1 + \log(\ell_i - \ell_{i-1})$ to a^{ℓ_i} (and make it the weight on the edge to the previous block). □

We now show that $\mathcal{G}_G$ can be transformed to $\mathcal{G}_{\text{TtoG}}$ without increasing the sum of weights of the edges. This is done by simple redirection of edges and changing their cost.

Lemma 18. *$\mathcal{G}_G$ can be transformed to $\mathcal{G}_{\text{TtoG}}$ without increasing the sum of weights of the edges.*

Proof. Fix a letter a, we show how to transform the subgraph of $\mathcal{G}_G$ induced by nodes labelled with blocks of a to the corresponding subgraph of $\mathcal{G}_{\text{TtoG}}$, without increasing the sum of weights.

Firstly, let us sort the nodes according to the increasing length of the blocks. For each node a^ℓ, if it has many edges, we delete all except one and then we redirect this edge to a^ℓ's direct predecessor (say a^k) and label it with a cost $1 + \log(\ell - k)$. This cannot increase the sum of weights of edges:

- deleting does not increase the sum of weights;
- if a_ℓ has an edge to ϵ with weight $1 + \log \ell$ then $1 + \log \ell \geq 1 + \log(\ell - k)$;
- otherwise it had an edge to some $k' \leq k$ with a weight $\ell - k'$. Then $1 + \log(\ell - k) \leq \ell - k \leq \ell - k'$, as claimed (note that $1 + \log x \leq x$ for $x \geq 1$).

Some blocks labelling nodes in $\mathcal{G}_G$ perhaps do not label the nodes in $\mathcal{G}_{\text{TtoG}}$. For such a block a^ℓ we remove its node a_ℓ and redirect its unique incoming edge to its predecessor, say $a_{\ell'}$, changing the weight appropriately. Since $1 + \log(x) + 1 + \log(y) \geq 1 + \log(x + y)$ when $x, y \geq 1$, we do not increase the total weight.

It is left to observe that if a node labelled with a^ℓ exists in $\mathcal{G}_{\text{TtoG}}$ then it also exists in $\mathcal{G}_G$, i.e. all blocks represented in TtoG occur in T. After RemCrBlocks there are no crossing blocks, see Lemma 12. So any maximal block in T (i.e. one represented by TtoG-based representation) is also a maximal block a^ℓ in some rule (after RemCrBlocks), say in X_i. But then this block is present in X_i also just before action of RemCrBlocks on X_i and so it is represented by G-based representation.

In this way we obtained a graph corresponding to the TtoG-based representation. □

Corollary 4. *The total cost of* TtoG-*representation is $\mathcal{O}(g \log N)$.*

Proof. By Lemma 18 it is enough to show this for the G-based representation, which holds by Corollary 3. □

4. Improved analysis

Intuitively, each "reasonable" grammar should have size $\mathcal{O}(|T|)$: application of a rule $X \to \alpha$ makes the current text longer by at least $|\alpha| - 1$, so the sum of all lengths of right-hand sides (so $|\alpha|$s) cannot be shorter than the input text. In some extreme cases this estimation might be better than $\mathcal{O}(g \log N)$ guaranteed by TtoG, thus TtoG should have an approximation guarantee $\mathcal{O}(\min(N, g \log N))$. This approach can be further improved: the trivial upper bound applies to any intermediate string obtained during TtoG and we can choose any of those estimations. We choose a specific point, where $|T| \approx g$. As a result, we divide the analysis of a computation of TtoG into two stages: the first one lasts while $|T| \geq g$ and then the second one begins. We separately estimate the cost of representation in the first stage, by $\mathcal{O}(g \log(N/g))$, and in the second, by $\mathcal{O}(g)$. In total this yields $\mathcal{O}(g + g \log(N/g))$; this matches the best known results for the smallest grammar problem [18,1,19] and is not worse than both $\mathcal{O}(g \log N)$ and $\mathcal{O}(g)$.

Theorem 2. *The* TtoG *runs in linear time and returns a grammar of size $\mathcal{O}\left(g + g \log\left(\frac{N}{g}\right)\right)$, where g is the size of the optimal grammar for the input text.*

Note that the time analysis was done already in Theorem 1, in the rest of this section we focus on the improved size analysis.

4.1. Outline

Firstly, in Section 4.2 we show that indeed any reasonable grammar for a text T has size $\mathcal{O}(T)$. This follows by simple calculation and shows that it is enough to calculate the cost of representation for the grammar when $|T| \geq g$. From Corollary 1 and Lemma 13 we know that those costs are covered by the issued credit and the additional representation cost for a-blocks. The analysis for credit is easy: since in each phase we introduce $\mathcal{O}(m) \leq \mathcal{O}(g)$ credit, it is enough to bound the number of phases and this follows from the fact that we shorten the text in each phase, see Lemma 7; this is done in

Section 4.3. On the other hand, the analysis of the representation cost for blocks is much more involved. The general outline remain as it was as in Section 3.6: we again use the G-based representation as a middle step, estimate its cost and compare it with the TtoG-based one. The difference is in the estimation of the G-based representation cost. We no longer can simply charge $\mathcal{O}(\log N)$ cost to a rule, we need a more subtle analysis. Instead of direct charging to a rule $X_i \to \alpha_i$, we associate the cost with some of the letters (of the original text) generated by X_i. To this end we 'mark' those letters and distinguish between different such markings. We ensure that such markings are disjoint, there are at most 2 of them per non-terminal and that the cost of representation is related to the total size of the markings, to be more precise, when we have markings of lengths $p_1, p_2, \ldots, p_k$ then the G-representation cost is $\mathcal{O}(\sum_{i=1}^{k} 1 + \log p_i)$. Then the estimation of the size of the whole grammar is just a matter of calculation. The markings and the analysis of representation cost using them is performed in Section 4.4. For technical reasons we also consider the cost of representation in the phase in which $|T|$ is reduced from more than g to smaller than g separately, the analysis is a simple combination of the case when $|T| > g$ and when $|T| < g$ and is done in Section 4.5. Wrapping up all estimations and giving the proof of Theorem 2 is done in Section 4.6. What is left is to describe the way we modify the markings to ensure their properties. This technical construction is presented separately in Section 4.7.

4.2. Linear bound

We begin with formalising the argument that any "reasonable" grammar has size $\mathcal{O}(|T|)$.

Lemma 19. *Let SLP G contain no production $X \to \alpha$ with $|\alpha| \leq 1$ and assume that every production is used in the derivation defined by G. Then $|G| \leq 2|T| - 1$.*

In particular, if at any point the letters created so-far by TtoG *have representation cost k and the remaining text is T then the final grammar for the input tree has size at most $k + 2|T| - 1$.*

Note that the grammar produced by TtoG clearly has the properties assumed by Lemma 19: we introduce new letters in place of substrings of length at least 2 and each of them is used in the derivation of the input text.

Proof. Assume that G has the properties from the lemma. An application of a rule $X_i \to \alpha_i$ to the current string increases its size by $|\alpha_i| - 1 \geq 1$ for each occurrence of X_i in the string derived so far. As we assume that each production is used in the derivation, each of $|\alpha_i| - 1 \geq 1$ is added at least once and so we get $\sum_{i=1}^{m}(|\alpha_i| - 1) \leq |T|$. Thus $\sum_{i=1}^{m} |\alpha_i| \leq |T| + m$ and so it is left to estimate m. As there are m productions and each application increases the size of the derived string by at least 1, and we start the derivation with a text of length 1, we get $m \leq |T| - 1$. Thus $\sum_{i=1}^{m} |\alpha_i| \leq |T| + m \leq 2|T| - 1$.

The second claim now easily follows: when the current string kept by TtoG is T then we can take as a whole grammar $X \to T$ together with the representation for the so-far created letters. □

In the following analysis we focus on the phase such that the text before it has length greater or equal to g and after it is smaller than g. Such phase exists: clearly $N \geq g$ (as we can take the grammar with T on the right-hand side) and so initially $|T| \geq g$ and in the end T is reduced to a single letter.

Lemma 20. *There is a phase in computation of* TtoG *such that at the beginning of the phase $|T| \geq g$ and at the end of the phase $|T| < g$.*

We separately estimate the cost of representation (i.e. issued credit and the cost of TtoG-based representation) up to the phase from Lemma 20, in this phase and after it. For the first two we show an upper bound of $\mathcal{O}(g + g\log(N/g))$, for the latter we use Lemma 19 to get an estimation $\mathcal{O}(g)$ on the representation cost.

4.3. Credit and pair compression when text is long

Lemma 21. *If at the beginning of the phase $|T| \geq g$ then $\mathcal{O}(g + g\log(N/g))$ credit was issued.*

Proof. Initial grammar G has at most g credit. The input text is of length N and the current one is of $t = |T|$ and so there were $\mathcal{O}(\log(N/t))$ phases, as in each phase the length of T drops by a constant factor, see Lemma 7. As $t \geq g$, we obtain a bound $\mathcal{O}(\log(N/g))$ on the number of phases. Due to Lemmata 10, 13, at most $\mathcal{O}(m) \leq \mathcal{O}(g)$ credit per phase is issued during the pair compression and block compression, so in total $\mathcal{O}(g + g\log(N/g))$ credit was issued. □

From Lemma 10 we know that the representation cost of letters introduced by pair compression is covered by the credit. Thus

Corollary 5. *Suppose that at the beginning of the phase $|T| \geq g$. Then the representation cost of letters introduced by pair compression till this phase is $\mathcal{O}(g + g\log(N/g))$.*

4.4. Cost of representing blocks when text is long

For the cost of representing blocks, we define the G-based and TtoG-based representations in the same way as previously. However, we slightly extend the notion: we consider those representations at any point of TtoG, not only at the end; this does not effect those notions in any way.

For both the G-based representation and the TtoG-based representation we again define graphs $\mathcal{G}_G$ and $\mathcal{G}_{\text{TtoG}}$ and by Lemma 16 the cost of G-based representation is $\Theta(w(\mathcal{G}_G))$ and by Lemma 17 the cost of TtoG-based representation is $\Theta(w(\mathcal{G}_{\text{TtoG}}))$. Then Lemma 18 shows that we can transform $\mathcal{G}_G$ to $\mathcal{G}_{\text{TtoG}}$ without increasing the sum of weights. Hence it is enough to show that the G-based representation cost is at most $\mathcal{O}(g + \log(N/g))$.

The G-based representation cost consists of some released credit and the cost of representing powers, see Lemma 16. The former was already addressed in Lemma 21 (the whole issued credit is $\mathcal{O}(g + g\log(N/g))$) and so it is enough to estimate the latter, i.e. the cost of representing powers.

The outline of the analysis is as follows: when a new power a^ℓ is represented, we mark some letters of the input text (and perhaps modify some other markings) those markings are associated with nonterminals and are named X_i-pre-power marking and X_i-in marking (which are defined in detail in Section 4.7). The markings satisfy the following conditions:

(M1) each marking marks at least 2 letters, no two markings mark the same letter;
(M2) for each X_i there is most one X_i-pre-power marking and at most one X_i-in marking;
(M3) when the substrings of length $p_1, p_2, \ldots, p_k$ are marked, then the so-far cost of representing the powers by G-based representation is $c\sum_{i=1}^{k}(1 + \log p_i)$ (for some fixed constant c).

We show that when we have a marking satisfying (M1)–(M3) then indeed the cost of representing blocks is $\mathcal{O}(g + g\log(N/g))$. The construction of the markings and the analysis of it is technical and does not affect further estimations of th grammar size, so it is moved to a separate Section 4.7.

Lemma 22. *If at the beginning of the phase $|T| \geq g$ then so far the cost of representing blocks by* TtoG *is $\mathcal{O}(g + g\log(N/g))$.*

Proof. The G-based representation cost consists of some released credit and the cost of representing powers, see Lemma 16. The former is bounded by $\mathcal{O}(g + g\log(N/g))$, see Lemma 21, in the following we estimate the cost of representing powers.

Using (M1)–(M3) the cost of representing powers (in G-based representation) can be upper-bounded by (a constant times):

$$k + \sum_{i=1}^{k} \log p_i, \text{ where } k \leq 2m \text{ and } \sum_{i=1}^{k} p_i \leq N \ . \tag{3a}$$

It is easy to show that (3a) is maximised for $k = 2m$ and each p_i equal to $N/2m$: clearly, the sum is maximised for $\sum_{i=1}^{k} p_i = N$. Then for a fixed k and $\sum_{i=1}^{k} p_i = N$ the sum $\sum_{i=1}^{k} \log p_i$ is maximised when all p_i are equal, which follows from the fact that $\log(x)$ is concave, hence we can set $p_i = \frac{N}{2m}$, arriving at

$$2k + 2k\log\left(\frac{N}{2k}\right) \ . \tag{3b}$$

The function $x + x\log(N/x)$ attains its maximum for $x = 2N/e$ and it monotonically increases on the interval $[0, 2N/e]$. If $g \leq N/e$ then (as $l \leq m \leq g$) we get that (3b) is at most

$$2g + 2g\log\left(\frac{N}{2g}\right) = \mathcal{O}\left(g + g\log\left(\frac{N}{g}\right)\right) \ . \tag{3c}$$

If $g \geq N/e$ then by Lemma 19 the trivial estimation for the size of the whole grammar returned by TtoG is $2N - 1 \leq 2eg - 1 = \mathcal{O}\left(g + g\log\left(\frac{N}{g}\right)\right)$, so also as claimed. □

4.5. Intermediate phase

We bounded the representation cost before the phase from Lemma 20 and after it, so it is left to estimate the cost within this phase.

Lemma 23. *The cost of representing letters in the phase from Lemma 20 is $\mathcal{O}(g + g\log(N/g))$.*

Proof. Let at the beginning of this phase T have length t_1 and after t_2, where $t_1 \geq g > t_2$.

The cost of representing letters introduced during the pair compression is covered by the released credit, see Lemma 10. There was at most $\mathcal{O}(g + g\log(N/g))$ credit in the grammar at the beginning of the phase, see Lemma 21, and during this phase at most $\mathcal{O}(g)$ credit was issued, see Lemma 10 and Lemma 13. So the total credit is $\mathcal{O}(g + g\log(N/g))$.

Consider the cost of representing blocks. This consists of representing blocks that are not powers, which is covered by the released credit, and the cost of representing powers. The former is already covered, as we already known that at most $\mathcal{O}(g + g\log(N/g))$ credit was issued till the end of this phase. Thus we consider only the cost of representing powers. As T at the end of the phase has t_2 letters, at most $2t_2$ letters representing blocks could be introduced in this phase (since at most two blocks can be merged into one letter by pair compression afterwards). Let $p_1, \ldots, p_k$ be the lengths of those powers. Then (see Lemma 3) the cost of representing those powers is proportional to

$$k + \sum_{i=1}^{k} \log p_i, \text{ where } k \le 2t_2 \text{ and } \sum_{i=1}^{k} p_i \le t_1 \ .$$

Using similar analysis as in the case of (3) it can be concluded that this is at most

$$2k + 2k\log\left(\frac{t_1}{2k}\right) \le 2k + 2k\log\left(\frac{N}{2k}\right)$$

with the equality following from $t_1 \le N$. Again, as in (3), if $g \le N/e$ then (recall that $k \le t_2 \le g$) this is

$$2k + 2k\log\left(\frac{N}{2k}\right) < 2g\log\left(\frac{N}{2g}\right) = \mathcal{O}\left(g\log\left(\frac{N}{g}\right)\right)$$

and if $g \ge N/e$ then Lemma 19 yields that the whole returned grammar has size $\mathcal{O}(N) = \mathcal{O}(g) = \mathcal{O}\left(g\log\left(\frac{N}{g}\right)\right)$. □

4.6. *Proof of Theorem 2*

Concerning the size of the returned grammar, consider the phase from Lemma 20. By Corollary 5 the cost of letters introduced before this phase by pair compression is $\mathcal{O}\left(g\log\left(\frac{N}{g}\right)\right)$. Similarly, the cost of representation of letters introduced by block compression before this phase is $\mathcal{O}\left(g\log\left(\frac{N}{g}\right)\right)$, see Lemma 22. The cost of representing letters introduced during this phase is also $\mathcal{O}\left(g\log\left(\frac{N}{g}\right)\right)$, by Lemma 23. Lastly, by Lemma 19, the cost of representing letters after this phase is $\mathcal{O}(g)$. Thus, in total, the whole representation cost is $\mathcal{O}\left(g\log\left(\frac{N}{g}\right)\right)$, as claimed.

4.7. *Markings' modification*

What is left to show is how to mark the letters and how to modify those markings so that (M1)–(M3) are preserved.

The idea of preserving (M1)–(M3) is as follows: if a new power of length ℓ is represented, this yields a cost $\mathcal{O}(1+\log\ell) = \mathcal{O}(\log\ell)$, see Lemma 16; we can choose c in (M3) so that this is at most $c\log\ell$ (as $\ell \ge 2$). Then either we mark new ℓ letters or we remove some marking of length ℓ' and mark $\ell \cdot \ell'$ letters, it is easy to see that in this way (M1)–(M3) is preserved (still, those details are repeated later).

Whenever we are to represent powers $a^{\ell_1}, a^{\ell_2}, \ldots,$ for each power a^ℓ, where $\ell > 1$, we find the right-most maximal block a^ℓ in T. Let X_i be the smallest nonterminal that derives (before RemCrBlocks) this right-most occurrence of maximal a^ℓ (clearly there is such a non-terminal, as X_m derives it). It is possible that this particular a^ℓ in X_is' rule was obtained as a concatenation of $\ell - k$ explicit letters to a^k (so, not as a power). In such a case we are lucky, as the representation of this a_ℓ is paid by the credit and we do not need to separately consider the cost of representing power a^ℓ. Otherwise the a^ℓ in this rule is obtained as a power and we mark some of the letters in the input that are 'derived' by this a^ℓ. The type of marking depends on the way this particular a^ℓ is 'derived': If one of the nonterminals in X_i's production was removed during RemCrBlocks, this marking is an *X_i-pre-power marking*. Otherwise, this marking is an *X_i-in marking*.

Lemma 24. *There is at most one X_i-pre-power marking.*

When X_i-in marking is created for a^ℓ then after the block compression X_i has two nonterminals inside its rule and between them there is exactly a_ℓ.

Proof. Concerning the X_i-pre-power marking, let a^ℓ be the first power that gets this marking. Then by the definition of the marking, afterwards in the rule for X_i there is at most one nonterminal. But this means that no power can be created in this rule later on, in particular, no new marking associated with X_i (pre-power marking or in-marking) can be created.

Suppose that a^ℓ was assigned an X_i-in marking, which as in the previous case means that the right-most occurrence of maximal block a^ℓ is generated by X_i but not by the nonterminals in the rule for X_i. Since a^ℓ is a power it is obtained in the rule as a concatenation of the a-prefix and the a-suffix popped from nonterminals in the rule for X_i. In particular this means that each nonterminal in the rule for X_i generates a part of this right-most occurrence of a^ℓ. If any of those nonterminals were removed during the block compression a^ℓ would be assigned an X_i-pre-power marking, which is not the case. So both those nonterminals remained in the rule. Hence after popping prefixes and suffixes, between those two nonterminals there is exactly a block a^ℓ, which is then replaced by a_ℓ, as promised. □

Consider the a^ℓ and the 'derived' substring w^ℓ of the *input text*. We show that if there are markings inside w^ℓ, they are all inside the last among those ws.

Lemma 25. *Let a^ℓ be an occurrence of a maximal block to be replaced with a_ℓ which 'generates' w^ℓ in the input text. If there is any marking within this w^ℓ then it is within the last among those ws.*

Proof. Consider any pre-existing marking within w^ℓ, say it was done when some b^k was replaced by b_k. As b_k is a single letter and a^ℓ derives it, each a derives at least one b_k. The marking was done inside the string generated by the right-most b_k (as we always put the marking within the rightmost occurrence of the string to be replaced). Clearly the right-most b_k is 'derived' by the right-most a within a^ℓ, so in particular it is inside the right-most w in this w^ℓ. So all markings within w^ℓ are in fact within the right-most w. □

We now demonstrate how to mark letters in the input text. Suppose that we replace a power a^ℓ, let us consider the right-most occurrence of this a^ℓ in T and the smallest X_i that generates this occurrence. This a^ℓ generates some w^ℓ in the input text. If there are no markings inside w^ℓ then we simply mark any ℓ letters within w^ℓ. In the other case, by Lemma 25 we know that all those markings are in fact in the last w. If any of them is the (unique) X_i-in marking, let us choose it. Otherwise choose any other marking. Let ℓ' denote the length of the chosen marking. Consider, whether this marking in w is unique or not

unique marking Then we remove it and mark arbitrary $\ell \cdot \ell'$ letters in w^ℓ; this is possible, as $|w| \geq \ell'$ and so $|w^\ell| \geq \ell \cdot \ell'$. Since $\log(\ell \cdot \ell') = \log \ell + \log \ell'$, the (M3) is preserved, as it is enough to account for the $1 + \log \ell \leq c \log \ell$ representation cost of a^ℓ as well as the $c \log \ell'$ cost associated with the previous marking of length ℓ'.

not unique Then $|w| \geq \ell' + 2$ (the 2 for the other markings, see (M1)). We remove the marking of length ℓ', let us calculate how many unmarked letters are in w^ℓ afterwards: in $w^{\ell-1}$ there are at least $(\ell - 1) \cdot (\ell' + 2)$ letters (by the Lemma 25: none of them marked) and in the last w there are at least ℓ' unmarked letters (from the marking that we removed):

$$\begin{aligned}(\ell - 1) \cdot (\ell' + 2) + \ell' &= (\ell\ell' + 2\ell - \ell' - 2) + \ell' \\ &= \ell\ell' + 2\ell - 2 \\ &> \ell\ell' \ .\end{aligned}$$

We mark those $\ell \cdot \ell'$ letters, as in the previous case, the associated $c \log(\ell\ell')$ is enough to pay for the cost.

There is one issue: it might be that we created an X_i-in marking while there already was one, violating (M2). However, we show that if there were such a marking, it was within w^ℓ (and so within the last w, by Lemma 25) and so we could choose it as the marking that was deleted when the new one was created. Consider the previous X_i-in marking. It was introduced for some power b^k, replaced by b_k that was a unique letter between the nonterminals in the rule for X_i, by Lemma 24. Consider the rightmost substring of the input text that is generated by the explicit letters between nonterminals in the rule for X_i. The operations performed on G cannot shorten this substring, in fact they often expand it. When b_k is created, this substring is generated by b_k, by Lemma 24. When a_ℓ is created, it is generated by a_ℓ, by Lemma 24, i.e. this is exactly w^ℓ. So in particular w^ℓ includes the marking for b_k.

This shows that (M1)–(M3) are preserved.

Acknowledgements

I would like to thank Paweł Gawrychowski for introducing me to the topic, for pointing out the relevant literature [14] and discussions; Markus Lohrey for suggesting the topic of this paper and bringing the idea of applying the recompression to the smallest grammar. Moreover, I would like to thank anonymous referees who pointed out many shortcomings as well as suggested improvements of the presentation.

This research was partially supported by NCN grant number 2011/01/D/ST6/07164, 2011–2015.

Appendix A. Sakamoto's algorithm [19]

In proof that bounds the number of introduced nonterminals [19, Theorem 2], it is first estimated that in one execution of the while loop for a factor f_i the introduced nonterminals occur in $f_1 f_2 \cdots f_{i-1}$, except perhaps a constant number of them. This argument follows from observation that f_i is compressed to $\alpha\beta\gamma$, where $|\alpha|$ and $|\gamma|$ are bounded by a constant and the earlier occurrence of the same string as f_i is compressed to $\alpha'\beta\gamma'$ (where also $|\alpha'|$ and $|\gamma'|$ are bounded by a constant). This is true, however, when α and γ represent nonterminals introduced by *repetition* procedure (i.e. they are blocks in the terminology used here) we need to take into the account also the additional nonterminals that are introduced for representation of those blocks. The estimation of $\mathcal{O}(1)$ is not enough, as in the worst case $\Omega(\log N)$ are needed to represent a single block of as. We do not see any easy patch to repair this flaw.

The improved analysis [19, Theorem 2], in which the number of nonterminals is bounded by $\mathcal{O}\left(g + \log\left(\frac{N}{g}\right)\right)$, has the same shortcoming.

References

[1] M. Charikar, E. Lehman, D. Liu, R. Panigrahy, M. Prabhakaran, A. Sahai, A. Shelat, The smallest grammar problem, IEEE Trans. Inform. Theory 51 (7) (2005) 2554–2576.

[2] P. Gawrychowski, Pattern matching in Lempel–Ziv compressed strings: fast, simple, and deterministic, in: C. Demetrescu, M.M. Halldórsson (Eds.), ESA, in: LNCS, vol. 6942, Springer, 2011.

[3] L. Gąsieniec, M. Karpiński, W. Plandowski, W. Rytter, Efficient algorithms for Lempel–Ziv encoding, in: R.G. Karlsson, A. Lingas (Eds.), SWAT, in: LNCS, vol. 1097, Springer, 1996.

[4] A. Jeż, Approximation of grammar-based compression via recompression, in: J. Fischer, P. Sanders (Eds.), CPM, in: LNCS, vol. 7922, Springer, 2013, full version available at http://arxiv.org/abs/1301.5842.

[5] A. Jeż, Recompression: a simple and powerful technique for word equations, in: N. Portier, T. Wilke (Eds.), STACS, in: LIPIcs, vol. 20, Schloss Dagstuhl–Leibniz Zentrum fuer Informatik, Dagstuhl, Germany, 2013, full version available at http://arxiv.org/abs/1203.3705, accepted to J. ACM, http://drops.dagstuhl.de/opus/volltexte/2013/3937.

[6] A. Jeż, The complexity of compressed membership problems for finite automata, Theory Comput. Syst. 55 (2014) 685–718, http://dx.doi.org/10.1007/s00224-013-9443-6.

[7] A. Jeż, One-variable word equations in linear time, Algorithmica (2015), http://dx.doi.org/10.1007/s00453-014-9931-3, in press.

[8] A. Jeż, Faster fully compressed pattern matching by recompression, ACM Trans. Algorithms 11 (3) (2015) 20:1–20:43, http://doi.acm.org/10.1145/2631920.

[9] A. Jeż, M. Lohrey, Approximation of smallest linear tree grammar, in: E.W. Mayr, N. Portier (Eds.), STACS, in: LIPIcs, vol. 25, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2014.

[10] M. Karpiński, W. Rytter, A. Shinohara, Pattern-matching for strings with short descriptions, in: CPM, 1995.

[11] J.C. Kieffer, E.-H. Yang, Sequential codes, lossless compression of individual sequences, and Kolmogorov complexity, IEEE Trans. Inform. Theory 42 (1) (1996) 29–39.

[12] N.J. Larsson, A. Moffat, Offline dictionary-based compression, in: Data Compression Conference, IEEE Computer Society, 1999.

[13] M. Lohrey, Algorithmics on SLP-compressed strings: a survey, Groups Complex. Cryptol. 4 (2) (2012) 241–299.

[14] K. Mehlhorn, R. Sundar, C. Uhrig, Maintaining dynamic sequences under equality tests in polylogarithmic time, Algorithmica 17 (2) (1997) 183–198.

[15] M. Mitzenmacher, E. Upfal, Probability and Computing: Randomized Algorithms and Probabilistic Analysis, Cambridge University Press, 2005.

[16] C.G. Nevill-Manning, I.H. Witten, Identifying hierarchical structure in sequences: a linear-time algorithm, J. Artificial Intelligence Res. 7 (1997) 67–82.

[17] W. Plandowski, Testing equivalence of morphisms on context-free languages, in: J. van Leeuwen (Ed.), ESA, in: LNCS, vol. 855, Springer, 1994.

[18] W. Rytter, Application of Lempel–Ziv factorization to the approximation of grammar-based compression, Theoret. Comput. Sci. 302 (1–3) (2003) 211–222.

[19] H. Sakamoto, A fully linear-time approximation algorithm for grammar-based compression, J. Discrete Algorithms 3 (2–4) (2005) 416–430.

[20] J.A. Storer, T.G. Szymanski, The macro model for data compression, in: R.J. Lipton, W.A. Burkhard, W.J. Savitch, E.P. Friedman, A.V. Aho (Eds.), STOC, ACM, 1978.

[21] A.C.-C. Yao, On the evaluation of powers, SIAM J. Comput. 5 (1) (1976) 100–103.