NP-Completeness of the Smallest 1-Level Grammar Problem

Johanne Müller Vistisen (12340226)
TU Wien, Austria
e12340226@tuwien.ac.at

June 20, 2024

Abstract

The Smallest 1-Level Grammar Problem focuses on identifying the most compact 1-level grammar capable of generating a given string, a challenge which is relevant in many research areas in need of space-efficient string compression. This specific variant of the problem considers grammars where every nonterminal, except the starting symbol, has a production rule that directly results in terminal strings. It was a long time open question whether the Smallest Grammar Problem was NP-hard, and initial proofs were limited to cases involving infinite alphabets. This paper examines the complexity of the Smallest 1-Level Grammar Problem for a finite alphabet. By reducing the well-known Vertex Cover Problem to the Smallest 1-Level Grammar Problem, we demonstrate its NP-hardness even for an alphabet of size five.

1 Introduction

Compressing a sequence of symbols, or a string, using a set of logical rules that allow for the reconstruction of the original string has numerous applications, particularly in data storage, transmission, and bioinformatics. This kind of compression can be seen as context-free grammar that only produces a single word. Producing such a grammar can be done in many different ways and this raises the question of whether it is possible to mathematically prove that a given set of rules achieves optimal compression, formally known as the Smallest Grammar Problem (SGP), which is the central focus of this seminar paper.

The Smallest Grammar Problem seeks to find the most compact context-free grammar that generates a given string, where the size of a grammar is the sum of the sizes of its rules. This problem is known to be NP-hard for unbounded alphabets, and it does not allow a polynomial-time approximation scheme unless $P = NP$, [3]. However, many practical applications involve finite alphabets, and it remained an open question whether the hardness results also apply to this more realistic scenario. In 2016, this gap was closed when it was shown that the Smallest Grammar Problem is indeed NP-hard even for grammars over finite alphabets [1].

This seminar paper shows a slightly different and maybe even surprising result by the same authors [2], namely that even for a 1-level grammar with a finite alphabet, finding the smallest grammar is NP-hard. This problem is known as the Smallest 1-Level Grammar Problem. This problem was addressed by the same researches who answered the complexity question of the Smallest Grammar Problem for finite alphabets, and this seminar paper follows and elaborates on the proof from their paper from 2021, [2].

2 Preliminaries

In this section, we will introduce some notation of this paper, explain what a context-free (1-level) grammar is, and introduce the Smallest 1-Level Grammar Problem. Most of the notation comes from the aforementioned article from 2021 [2].

Notation In this paper $\mathbb{N}$ denotes the set of natural numbers, $\{0, 1, 2, \dots\}$. We measure the size of a set, A, by $|A|$. The symbol Σ denotes a finite alphabet, where $|\Sigma|$ is the size of this alphabet. The empty word is represented by ε and has size 0. We can create *words* from our finite alphabet, which are finite sequences of symbols from Σ. The set of non-empty words is denoted Σ^+, and we denote the set of all words $\Sigma^* = \Sigma^+ \cup \{\varepsilon\}$. The size of a word, w, is written as $|w|$, and the notation $|w|_a$ is the number of times that the symbol a occurs in w.

Concatenating two words, w_1 and w_2, together (meaning creating the word $w_1 w_2$) is done by the $\cdot$ operator, such that we have $w_1 \cdot w_2 = w_1 w_2$, and $w_1^r = \underbrace{w_1 \cdot w_1 \cdot \ldots \cdot w_1}_{r \text{ times}} = w_1 w_1 \dots w_1$ for $r \in \mathbb{N}$.

A context-free grammar, G is defined as $G = (N, \Sigma, R, S)$. N denotes the nonterminals of the grammar, Σ is the terminal alphabet, R is the set of rules of G. A rule is a pair $(A, w) \in R$ where given an occurrence of A in a word, we replace this with w. The rule can also be written as $A \to w$. Lastly, S is the starting symbol of the grammar; sometimes, instead of S, we write the grammar's axiom, ax. The axiom of the grammar is the right-hand side of the rule of S e.i. $S \to \text{ax}$.

We will look mostly at singleton grammars, which are context-free grammars where every nonterminal has precisely one production rule. If we draw a directed graph every time a production rule A goes to another nonterminal B, then this graph has no cycles. Said in another way; the relation $\{(A, B) | (A, w) \in R, |w|_B \geq 1\}$ is acyclic.

The size of a grammar is the size of the axiom of the grammar and the sum of the sizes of all the right sides of the grammar rules:

$$|G| = |\text{ax}| + \sum_{A \in R} |\mathfrak{D}(A)|.$$

When we have a word w and a grammar that can produce this word from a set of rules and an axiom, then we say that the derivative of the grammar is w, or more formally, $\mathfrak{D}(G) = w$. Taking $\mathfrak{D}(E)$ of some word E (regarding some grammar G), which can consist of both terminals and nonterminals, means using the rules of the nonterminals in E until everything that remains are terminal symbols. Taking $\mathfrak{D}(G)$ means taking $\mathfrak{D}(\text{ax})$, which should always lead to the word w in our singleton grammars. A way to visualise $\mathfrak{D}(\text{ax})$ is to build a tree where each level of the tree is one evaluation of a rule. All the leaves are then terminal characters, and we find the word w by concatenating the leaves from left to right. In a 1-level singleton grammar, every nonterminal, except for the starting symbol, has a rule that leads to a word for terminals. Said in another way, the derivation tree of G only has one level.

The smallest (1-level) grammar problem is, as mentioned in the introduction, finding out whether, for a given word w, there exists a 1-level grammar G of size k or smaller for some $k \in \mathbb{N}$. We can formulate the problem a bit more formally as:

Smallest 1-Level Grammar Problem (1-SGP)

Input: A word w and a $k \in \mathbb{N}$

Question: Does there exist a level-1 grammar G with $\mathfrak{D}(G) = w$ and $|G| \leq k$?

We can easily observe that this problem is in NP, since just guessing a grammar G and checking if $\mathfrak{D}(G) = w$ and $|G| \leq k$ can be done in polynomial time.

3 NP-hardness of 1-SGP

We will show that 1-SGP is NP-hard by reducing the well-known NP-hard Vertex Cover Problem to 1-SGP. Vertex Cover Problem is defined as follows

Vertex Cover Problem (VC)

Input: A graph $\mathcal{G} = (V, E)$, where and a $k \in \mathbb{N}$

Question: Does there exist a $C \subseteq V$, such that for every $(u, v) \in E$ it holds that $\{u, v\} \cap C \neq \emptyset$ and $|C| \leq k$?

The reduction goes as follows; given a graph $\mathcal{G} = (V, E)$, where $n := |V|$ and $m := |E|$, we wish to solve the Vertex Cover Problem for k. To do so we make the following constructions[1]. We define an alphabet of five symbols, $\Sigma = \{\mathtt{a}, \mathtt{b}, \diamond, \star, \#\}$ and the abbreviation $[\diamond] = \diamond^{n^3}$. Then for each vertex $v_i \in V$ we create a word $\overline{v_i} \in \{\mathtt{a}, \mathtt{b}\}^{\lceil \log n \rceil}$ such that no two vertices corresponds to the same word ($v_i = v_j$ if and only if $i = j$)[2]. Finally, we construct a word, w:

$$w = \prod_{i=1}^{n} (\#\overline{v_i}[\diamond]\overline{v_i}\#[\diamond])^{2\lceil \log(n) \rceil + 3} \prod_{i=1}^{n} (\#\overline{v_i}\#[\diamond])^{\lceil \log(n) \rceil + 1} \prod_{i=1}^{m} \left(\#\overline{v_{j_{2i-1}}}\#\overline{v_{j_{2i}}}\#[\diamond]\right)^2 \star [\diamond]^{n^3}. \tag{1}$$

Now, to show that the reduction from VC to 1-SGP works, we will show that we can find a vertex cover of size at most k if and only if we can find the smallest 1-level grammar of size at most $k' = 13n\lceil \log(n) \rceil 17n + k + 6m + 1 + 2n^3$, where n is the number of vertices and m is the number of edges in $\mathcal{G}$.

If there exists a vertex cover of size k, then there exists a 1-level grammar of size at most k'

Lemma 1. *If there exists a vertex cover of $\mathcal{G}$ of size k then there exists a grammar G with $\mathfrak{D}(G) = w$ and $|G| = k' = 13n\lceil \log(n) \rceil 17n + k + 6m + 1 + 2n^3$.*

Proof. We are given a vertex cover $\Gamma \subseteq V$ of the graph $\mathcal{G}$. Now we construct the following grammar, $G = (n, \Sigma, R, \mathrm{ax})$;

[1] We will, without loss of generality, assume through this that $n := |V| \geq 40$. This can be done since we can just assume that we have stored all the answers to the finite number of problems where $n < 40$.

[2] This can be done in different ways, e.g., by taking the binary representation of i using $\mathtt{a}$ and $\mathtt{b}$ as the binary alphabet.

$$N = \left\{ D, \overleftarrow{V_i}, \overrightarrow{V_i}, \overleftrightarrow{V_j}, \mid 1 \leq i \leq n, v_j \in \Gamma \right\}$$
$$R = \{S \to u, D \to [\diamond]\} \cup \left\{ \overleftarrow{V}_i \to \#\overline{v}_i, \overrightarrow{V}_i \to \overline{v}_i\# \mid 1 \leq i \leq n \right\} \cup \left\{ \overleftrightarrow{V_j} \to \#\overline{v_j}\# \mid v_j \in \Gamma \right\}$$
$$\text{ax} = \prod_{i=1}^{n} \left(\overleftarrow{V_i} D \overrightarrow{V_i} D \right)^{2\lceil \log(n) \rceil + 3} \prod_{i=1}^{n} (y_i D)^{\lceil \log(n) \rceil + 1} \prod_{i=1}^{m} (z_i D)^2 \star D^{n^3}$$

where for every $i \in \{1, 2, \ldots, n\}$ we have $y_i = \overleftrightarrow{V_i}$ if $v_i \in \Gamma$ and $y_i = \overleftarrow{V_i}\#$ otherwise. Furthermore for every $i \in \{1, 2, \ldots, m\}$, $z_i = \overleftrightarrow{V}_{j_{2i-1}} \overrightarrow{V}_{j_{2i-1}}$ if $v_{j_{2i-1}} \in \Gamma$ and $z_i = \overleftarrow{V}_{j_{2i-1}} \overleftrightarrow{V}_{j_{2i}}$ if $v_{j_{2i-1}} \notin \Gamma$. Since the pair of vertices $(v_{j_{2i-1}}, v_{j_{2i}})$ are connected by an edge, at least one of them has to be in the vertex cover Γ, since otherwise the edge $(v_{j_{2i-1}}, v_{j_{2i}})$ is uncovered. It is easy to see that this grammar is a 1-level grammar since all the nonterminals of N have rules that end in terminal characters.

Now, we move on to checking whether $\mathfrak{D}(G) = w$. Replacing all the nonterminals in the axiom with the corresponding rules results in

$$\prod_{i=1}^{n} (\#\overline{v_i}[\diamond]\overline{v_i}\#[\diamond])^{2\lceil \log(n) \rceil + 3} \prod_{i=1}^{n} (y_i[\diamond])^{\lceil \log(n) \rceil + 1} \prod_{i=1}^{m} (z_i[\diamond])^2 \star [\diamond]^{n^3}.$$

The only thing left to look at is the y_is and z_is. But since y_i is either $\overleftrightarrow{V_i}$ which goes to $\#\overline{v_i}\#$ or $\overleftarrow{V_i}\#$ which goes to $\#\overline{v_i}\#$, y_i is going to be replaced by the same word either way. It is quite similar to z_i, which is either $\overleftrightarrow{V}_{j_{2i-1}} \overrightarrow{V}_{j_{2i}}$ or $\overleftarrow{V}_{j_{2i-1}} \overleftrightarrow{V}_{j_{2i}}$, both resulting in $\#\overline{v_{j_{2i-1}}}\#$.

The only thing left to assert is whether $|G|$ has the size we claimed it had. First, we look at the size of the rules. The rule $\overleftarrow{V}_i \to \#\overline{v_i}$ has size $\lceil \log(n) \rceil + 1$ (because of the way $\overline{v_i}$ previously were defined), and it is the same for the rule $\overrightarrow{V}_i \to \overline{v_i}\#$, which also has size $\lceil \log(n) \rceil + 1$. Since we have one of each of these two rule types for each vertex, $v_i \in V$, those two rules (for all vertices combined) in total have size $2n\,(\lceil \log(n) \rceil + 1) = 2n\lceil \log(n) \rceil + 2n$. For each vertex, v_j, in the vertex cover there is also a rule $\overleftrightarrow{V}_j$ that goes to $\#\overline{v}_j\#$, since the vertex cover has size $|\Gamma| = k$ and the rule has size $\lceil \log(n) \rceil$, this contributes with $k(\lceil \log(n) \rceil + 2) = k\lceil \log(n) \rceil + 2k$. The rule $D \to [\diamond]$ has size n^3. Hence, the rules, when all added together, have size

$$2n\lceil \log(n) \rceil + 2n + k\lceil \log(n) \rceil + 2k + n^3.$$

We find the size of the axiom by counting the number of symbols (terminals and nonterminals) that occur in the axiom. From the first product of ax we get $n \cdot 4(2\lceil \log(n) \rceil + 3)$ symbols. From the next product we get $(3n - k)(\lceil \log(n) \rceil + 1)$ symbols because y_i uses 3 symbols unless $v_i \in \Gamma$ in that case y_i only consists of 2 symbols. The last product has $m(3 \cdot 2) = 6m$ symbols, the $\star$ is a single symbol and D^{n^3} consists of n^3 symbols. Thus, the size of the axiom is

$$4n(2\lceil \log(n) \rceil + 3) + (3n - k)(\lceil \log(n) \rceil + 1) + 6m + 1 + n^3.$$

Abbreviated a bit and added to the size of the rules, we end up with a grammar of the wished

size,

$$\begin{aligned}
&4n(2\lceil\log(n)\rceil+3)+(3n-k)(\lceil\log(n)\rceil+1)+6m+1+n^3\\
&\quad+2n\lceil\log(n)\rceil+2n+k\lceil\log(n)\rceil+2k+n^3\\
&=11n\lceil\log(n)\rceil-k\lceil\log(n)\rceil+15n-k+6m+1+n^3.\\
&\quad+2n\lceil\log(n)\rceil+2n+k\lceil\log(n)\rceil+2k+n^3\\
&=13n\lceil\log(n)\rceil+17n+k+6m+1+2n^3
\end{aligned}$$

This concludes the proof. □

If there exists a 1-level grammar of size k', then there exists a vertex cover of size at most k. The word w continues to be defined as previously and the same for the graph $\mathcal{G}=(V,E)$. This part of the reduction will establish that there indeed exists a vertex cover of size at most k if a grammar of size $k'=13n\lceil\log(n)\rceil+17n+k+6m+1+2n^3$ exists.

Lemma 2. *Let G be a 1-level grammar, $G=(N,\Sigma,R,\mathrm{ax})$. Then $|G|\geq 2\sqrt{|\mathfrak{D}(G)|}$.*

Proof. Let $n:=|\mathfrak{D}(G)|$ and let ax be the axiom of G. We then chose a rule $A\to u$ where $|u|$ is the biggest of all the rules in R. The number of symbols in the axioms times the size of u will then always be bigger than or equal to $|\mathfrak{D}(G)|$ since $A\to u$ is the rule of maximum length and the size of $\mathfrak{D}(G)$ comes from going through all symbols of the axiom and summing their sizes. Hence we get

$$|\mathrm{ax}||u|\geq|\mathfrak{D}(G)|.$$

Now we use the small trick that $x+y\geq 2\sqrt{xy}$, for all $y,x\geq 0$. Therefore

$$|\mathrm{ax}|+|u|\geq 2\sqrt{|\mathrm{ax}||u|}.$$

Now, we use the fact that $|G|$ is the size of the axiom and the size of all the rules of G put together. This means that $|G|$ is greater or equal to the size of the axiom and just the size of a single rule;

$$|G|\geq|\mathrm{ax}|+|u|\geq 2\sqrt{|\mathrm{ax}||u|}\geq 2\sqrt{n}=2\sqrt{|\mathfrak{D}(G)|}.$$

We have now shown that the compression of G is at most quadratic. □

Lemma 3. *Let $G=(N,\Sigma,R,\mathrm{ax})$ be a smallest 1-level grammar with $\mathfrak{D}(G)=w$ and $|G|=k'$ then $|G|<\frac{5n^3}{2}$.*

Proof. We begin with a couple of observations. Since k is going to be the size of a vertex cover of $\mathcal{G}=(V,E)$, then obviously $|V|=n\geq k$. The number of edges in a simple graph can be upper bounded by $m\leq n(n-1)=n^2-n$ if we assume that there are no self-loops or double edges in $\mathcal{G}$, very specifically $m+1\leq n^2-n+1<n^2$. From our assumptions, we have

$$|G|\leq 13n\lceil\log(n)\rceil+17n+k+6m+1+2n^3 \tag{2}$$

Since we continue to assume that $n\geq 40$, we then can establish the following (note that we have not included $2n^3$ from $|G|$);

$$\begin{aligned}
13n\lceil\log(n)\rceil+17n+k+6m&<13n^2+17n+n+6n^2 \qquad (3)\\
&=19n^2+18n\\
&<20n^2.
\end{aligned}$$

From $20n^2$ we can get to $\frac{40}{2}n^2 \leq \frac{n}{2}n^2 = \frac{n^3}{2}$ still by using the fact that $n \geq 40$. If we add $\frac{n^3}{2}$ to the $2n^3$ we did not include in (3) we obtain a strict upper bound for the size of G namely,

$$|G| < \frac{n^3}{2} + 2n^3 = \frac{5n^3}{2}. \tag{4}$$

□

Lemma 4. *Let $G = (N, \Sigma, R, \text{ax})$ be a smallest 1-level grammar of size k' with $\mathfrak{D}(G) = w$ then $2n^3 \leq |G|$.*

Proof. We begin with an observation: The symbol splits the axiom of G in two. This is because $\star$ only occurs once in w and therefore splits w into two parts with non-overlapping rules (because it does not make sense to make a rule for a symbol only occurring once or include this symbol in any other rules). Therefore we have that the axiom of G can actually be written as $u \star u'$, where $\mathfrak{D}(u') = [\diamond]^{n^3}$.

We move on to the lower bound of $|G|$. As just mentioned the axiom of G has the form $u \star u'$ and $\mathfrak{D}(u') = [\diamond]^{n^3}$. If we concern ourselves only with the grammar connected to u', let us denote it G' then this grammar have $\mathfrak{D}(u') = [\diamond]^{n^3}$ which is a word of size n^6 because $[\diamond]^{n^3} = (\diamond^{n^3})^{n^3} = (\underbrace{\diamond\diamond\ldots\diamond}_{n^3 \text{ times}})^{n^3}$. Consequently $\mathfrak{D}(u')$ is going to be a word of $n^3 \cdot n^3 = n^6$ symbols. But from Lemma 2 we have that

$$|G'| \geq 2\sqrt{\mathfrak{D}(u')} = 2\sqrt{n^6} = 2n^3.$$

Since $|G| \geq |G'|$ we now have

$$2n^3 \leq |G|.$$

□

Lemma 5. *Let $G = (N, \Sigma, R, \text{ax})$ be a smallest 1-level grammar with $\mathfrak{D}(G) = w$, then there is a $D \in N$ where $D = [\diamond]$ and for every other rule $A \to x$ in R, $|x|_\diamond = 0$.*

Proof. Earlier, we made the observation that the axiom of G must be of the form $u \star u'$. We will begin by looking at the subword u' and how to optimally produce it. Let us first assume that there is a rule $A \to \diamond^l$ with $l > n^3$. This rule can only be used on the suffix of w since all other occurrences of $\diamond$ are strictly smaller than $[\diamond]^{n^3}$. The question is, then, what is the most optimal rule to compress the subword $[\diamond]^{n^3}$? If we pick $A \to [\diamond]$, then the axiom of the grammar G then becomes $u \star A^{n^3}$ (recall that we just showed that we could separate the axiom of G into $u \star u'$). Applying Lemma 2 on the sub-grammar G' with derivative $\mathfrak{D}(G')$ we get that

$$|G'| \geq 2\sqrt{|\mathfrak{D}(G')|} = 2\sqrt{|[\diamond]^{n^3}|} = 2n^3$$

And since the axiom $|\text{ax}| + |\text{right side of } A| = |A^{n^3}| + |[\diamond]| = 2n^3$ the rule $A \to [\diamond]$ compresses $[\diamond]^{n^3}$ optimally because it results in $|G'| = 2n^3$. This means that there at least does not exist a rule $A \to \diamond^l$ with $l > n^3$.

We still do not know what a rule that deals with $[\diamond]$ looks like, but since the size of G is strictly smaller than $3n^3$, we can only have at most two nonterminals that produce some factor of $[\diamond]$. The reason is that $[\diamond]$ occurs n^3 times and, therefore, a rule, let us denote this rule δ, that deals with some factor of $[\diamond]$ will take up some number of spaces in the axiom $|\delta^{n^3}| = |\delta| \cdot n^3$. Which means that $\delta \leq 2$ in order to have $|G| = |\text{ax}| + |\text{rules of } G| \leq 3n^3$. But if two nonterminals are used to construct $[\diamond]$, then one of these nonterminals must have a right side that is larger or equal to $\frac{|[\diamond]|}{2}$. Let us now construct such a rule and call it $B \to v$ with $|v| \geq \frac{|[\diamond]|}{2} = \frac{n^3}{2}$. Now we have the following claim:

Claim 1. There cannot be a rule in G that contains a symbol from $\Sigma \setminus \{\diamond\}$ with a right-side that has size more than or equal to $\frac{n^3}{2}$.

Proof of claim 1. To prove this claim, we begin by noticing that a rule, here denoted R_1, with a symbol $\Sigma \setminus \{\diamond\}$ will not be a rule in G'. This means that the size of G will at least be

$$|G| \geq |G'| + |R_1| = 2n^3 + \frac{n^3}{2} = \frac{5n^3}{2}.$$

But this is a contradiction to Lemma 3 which stated that $|G| < \frac{5n^3}{2}$. This proves the claim. (Claim 1) □

We will now show why we can replace every rule $B \to v$ occurrence in G' with the rule $D \to [\diamond]$. Let us assume that the $|v|$ is the largest among all the right sides of the rules in G'. Let us denote $|v| = n^3 - t$. If we now look at u' (from the axiom of G, $\text{ax} = u \star u'$). The size of u' is

$$|u'| \geq \frac{n^6}{n^3 - t} > \frac{n^6 - t^2}{n^3 - t} = \frac{(n^3 + t)(n^3 - t)}{n^3 - t} = n^3 + t.$$

but since the size of the axiom would be n^3 if we used the rule $D \to [\diamond]$ we can remove every occurrence of B in u' without increasing the grammar.

So we can replace the rule of B everywhere in u', but it might have been the case that the rule is used in u, which means that what is left of the proof is to show that D is actually the only rule which is used to replace $[\diamond]$. Let us assume that there is some other sequence of terminals and nonterminals such that their rules produce $[\diamond]$ of u (that is, the occurrences of $[\diamond]$ on the left side of $\star$). Theses symbols will be of the following form; $E_1 C_1 \dots C_p E_2$, and have the derivative $\mathfrak{D}(E_1 C_1 \dots C_p E_2) = x \diamond^q y$. With this notation, we have $E_1 \to x\diamond^q$ with $q \geq 1$ or $E \to \varepsilon$, and similarly $E_2 \to \diamond^r y$, for $y \geq 1$. Now the trick will be to exchange this long expression of symbols $E_1 C_1 \dots C_p E_2$ to $E_1' D E_2'$ and showing that this does not increase the size of the grammar. The way E_1' and E_2' are defined is the following; $E_1' \to x$ if $E_1 \to x\diamond^q$ and if $E_1 = \varepsilon$ then $E_1' = \varepsilon$, a very similar definition applies to E_2'; $E_2' \to y$ if $E_2 \to \diamond^r y$ and if $E_2 = \varepsilon$ then $E_2' = \varepsilon$.

First, we show that in $E_1 C_1 \dots C_p E_2$ p must be greater or equal to 1. This is because if $p = 0$ then $\mathfrak{D}(E_1 E_2) = x[\diamond]y$, but since there only are two rules, E_1 and E_2, that produce this derivative, one of these rules must at least have size $\frac{n^3}{2}$ because $x[\diamond]y \geq \frac{n^3}{2}$. In Claim 1, we showed that it is impossible for a rule to have a size greater than $\frac{n^3}{2}$ and contain any symbol from Σ other than $\diamond$. Hence, we know that $p \geq 1$. We now see that $|E_1| \geq |E_1'|$ and $|E_2| \geq |E_2'|$ by their definitions, and since $p \geq 1$ we have that $|C_1 \dots C_p| \geq |D|$, which means that we can replace $E_1 C_1 \dots C_p E_2$ with $E_1' D E_2'$ without increasing the size of the grammar. We can now conclude that every occurrence of $[\diamond]$ in w is produced by the rule D, and even if there were more rules that produced the symbol $\diamond$, they would be superfluous, and we could remove them. This concludes the proof of the Lemma. □

Lemma 6. *If there exists a 1-level grammar G with $\mathfrak{D}(G) = w$ and $|G| \leq 13n\lceil\log(n)\rceil + 17n + k + 6m + 1 + 2n^3$, then there exists a size k vertex cover of $\mathcal{G}$.*

Proof. We can bound the size of G both from above and from below the following way because of Lemma 3 and Lemma 4, which is

$$2n^3 \leq |G| < \frac{5n^3}{2} < 3n^3.$$

Having established the bounds of $|G|$, we establish the form that the axiom of G must eventually take. Since G produce the word w as defined earlier, the axiom must have the form

$$\text{ax} = \prod_{i=1}^{n} \left(\alpha_i[\diamond]\alpha_i'[\diamond]\right)^{2\lceil\log(n)\rceil+3} \prod_{i=1}^{n} \left(\beta_i[\diamond]\right)^{\lceil\log(n)\rceil+1} \prod_{i=1}^{m} \left(\gamma_i[\diamond]\right)^2 \star [\diamond]^{n^3}. \tag{5}$$

From Lemma 5 we know that all occurrence of $[\diamond]$ can be produced by $D \to [\diamond]$ and therefore the axiom becomes

$$\text{ax} = \prod_{i=1}^{n} \left(\alpha_i D\alpha_i' D\right)^{2\lceil\log(n)\rceil+3} \prod_{i=1}^{n} \left(\beta_i D\right)^{\lceil\log(n)\rceil+1} \prod_{i=1}^{m} \left(\gamma_i D\right)^2 \star D^{n^3}.$$

In the following claims, we will show α_i, α_i' and β_i (for $1 \le i \le n$), and γ_i (for $1 \le i \le m$) can be replaced in the axiom by symbols that we have seen in Lemma 1 without increasing the size of the grammar.

Claim 2. For every $1 \le i \le n$ we have $\alpha_i = \overleftarrow{V_i}$ and $\alpha_i' = \overrightarrow{V_i}$, where $\overleftarrow{V_i}$ and $\overrightarrow{V_i}$ are nonterminals with rules $\overleftarrow{V_i} \to \#\overline{v_i}$ and $\overrightarrow{V_i} \to \overline{v_i}\#$.

Proof of claim 2. For every i, $1 \le i \le n$, we know that $\mathfrak{D}(\alpha_i) = \#\overline{v_i}$. If $|\alpha_i| = 1$, then α_i would have to be a single symbol, namely a nonterminal, with a rule that resulted in $\#\overline{v_i}$. This nonterminal could, for instance, be $\overleftarrow{V}_i \to \#\overline{v_i}$. If then for some reason $|\alpha_i| \ge 2$ we could substitute this α_i by $\overleftarrow{V}_i$ with rule $\overleftarrow{V}_i \to \#\overline{v_i}$. It would not increase the grammar to substitute α_i by $\overleftarrow{V}_i$ because even though the new rule would increase the grammar by $\lceil\log(n)\rceil+1$, it would also decrease the axiom by $2\lceil\log(n)\rceil+3$ because $|\overleftarrow{V}_i| = 1$. Hence, the grammar is not increased by substituting all α_i with size more than 1. One can use a similar argument to show that α_i' with a size more than 1 can be replaced by $\overrightarrow{V}_i$ with rule $\overrightarrow{V}_i \to \overline{v_i}\#$. This shows the claim, and we conclude that the grammar G has the two nonterminals $\overleftarrow{V}_i$ and $\overrightarrow{V}_i$ with corresponding rules $\overleftarrow{V}_i \to \#\overline{v_i}$ and $\overrightarrow{V}_i \to \overline{v_i}\#$. (Claim 2) □

Claim 3. For every i, $1 \le i \le n$, $\beta_i = \overleftarrow{V}_i\#$ or $\beta_i = \overleftrightarrow{V}_i$ where $\overleftarrow{V}_i$ and $\overleftrightarrow{V}_i$ are nonterminals with rules $\overleftarrow{V}_i \to \#\overline{v_i}$ and $\overleftrightarrow{V}_i \to \#\overline{v_i}\#$.

Proof of claim 3. To prove this claim, we look at two different cases $|\beta_i| = 1$ and $|\beta_i| \ge 2$ for i, $1 \le i \le n$. If $|\beta_i| = 1$ then there is some rule with a nonterminal $C \to \#\overline{v_i}\#$. We could replace C with $\overleftrightarrow{V}_i$ and then have the rule $\overleftrightarrow{V}_i \to \#\overline{v_i}\#$. Otherwise if $|\beta_i| \ge 2$ for some i, $1 \le i \le n$, then we could replace those β_i with $\overleftarrow{V}_i\#$. This would not increase the grammar, since the $|\beta_i| \ge |\overleftarrow{V}_i\#| = 2$ and the derivative of rule remains $\#\overline{v_i}\#$ for every i, $1 \le i \le n$. In conclusion, for every i, $1 \le i \le n$, either $\beta_i = \overleftrightarrow{V}_i$ or $\beta_i = \overleftarrow{V}_i\#$. (Claim 3) □

Now the last unknown symbols of the axiom are γ_i for every i for $1 \le i \le m$. We recall that $\mathfrak{D}(\gamma_i) = \#\overline{v_{j_{2i-1}}}\#\overline{v_{j_{2i}}}\#$ for every j, $1 \le j \le m$. We will look at three distinct cases for the size of γ_i; $|\gamma_i| = 1$, $|\gamma_i| = 2$, and $|\gamma_i| \ge 3$.

With $|\gamma_i| = 1$, we must have a nonterminal with a rule of the following form $E \to \#\overline{v_{j_{2i-1}}}\#\overline{v_{j_{2i}}}\#$. This rule has a right side of size $2\lceil\log(n)\rceil+3$, and it will add 2 to the size of the axiom (because we have γ_i^2). If we instead look at $\overleftarrow{V}_{j_{2i-1}}\overrightarrow{V}_{j_{2i}}\#$, the derivative of this will be of size $2\lceil\log(n)\rceil+3$ and it will add 6 to the axiom. But since $\#\overline{v_{j_{2i-1}}}\#\overline{v_{j_{2i}}}\#$ only occurs in w in γ_i, we can conclude that removing the rule $E \to \#\overline{v_{j_{2i-1}}}\#\overline{v_{j_{2i}}}\#$ will decrease the grammar by $2\lceil\log(n)\rceil+3$ and since we already have introduced the nonterminals $\overleftarrow{V}_{j_{2i-1}}$, $\overrightarrow{V}_{j_{2i}}$, and their rules we only increase the axiom with 4 by using these rules and since $4 \le 2\lceil\log(n)\rceil+3$ we still have not increased the size of the grammar.

If $|\gamma_i| = 2$ then we know that there must be two nonterminals, E_1 and E_2 such that $\mathfrak{D}(E_1E_2) = \#\overline{v_{j_{2i-1}}}\#\overline{v_{j_{2i}}}\#$, and either $E_1 \to \#\overline{v_{j_{2i-1}}}\#x$ or $E_2 \to x\#\overline{v_{j_{2i}}}\#$. Let us begin by looking at the case of $E_1 \to \#\overline{v_{j_{2i-1}}}\#x$ then we can change the rule E_1 so we remove the x and the rule becomes $E_1 \to \#\overline{v_{j_{2i-1}}}\#$ which is equivalent to the rule $\overleftrightarrow{V}_{j_{2i-1}}$ and to obtain the rest of $\#\overline{v_{j_{2i-1}}}\#\overline{v_{j_{2i}}}\#$ we make $E_2 = \overrightarrow{V}_{j_{2i}}$. We do all this without increasing the grammar since the size of the right-hand side of the rule remains the same, and the fact that we have not introduced new rules. The case of $E_2 \to x\#\overline{v_{j_{2i}}}\#$ is very similar and there we end up with replacing E_2 with $\overleftrightarrow{V}_{j_{2i}}$ and E_1 with $\overleftarrow{V}_i$

Finally if $|\gamma_i| \geq 3$ we can replace every occurrence of γ_i with $\overleftarrow{V}_{j_{2i-1}}\overleftarrow{V}_{j_{2i}}\#$ without increasing the size of the grammar, since the size of $\overleftarrow{V}_{j_{2i-1}}\overleftarrow{V}_{j_{2i}}\#$ is 3 and the derivative of γ_i remains unchanged.

Based on the previous statements as well as Claim 2 and Claim 3, we can conclude that the rules of G are the following: $D \to [\diamond]$, $\overleftarrow{V}_i \to \#\overline{v_i}$ and $\overrightarrow{V}_i \to \overline{v_i}\#$ for $1 \leq i \leq n$ and lastly for all $i \in \mathfrak{J}$ for some $\mathfrak{J} \subseteq \{1, 2, \ldots, n\}$ we have the rule $\overleftrightarrow{V}_i \to \#\overline{v_i}\#$. We denote $|\mathfrak{J}| = l$ and the corresponding vertex set is defined the following way $\mathcal{V} = \{v_i \mid i \in \mathfrak{J}\}$. Finally, we define t as the number of edges covered by $\mathcal{V}$. Now, the axiom looks the following way

$$\mathrm{ax} = \prod_{i=1}^{n} \left(\overleftarrow{V}_i D \overrightarrow{V}_i D\right)^{2\lceil \log(n)\rceil+3} \prod_{i=1}^{n} (y_i D)^{\lceil \log(n)\rceil+1} \prod_{i=1}^{m} (z_i D)^2 \star D^{n^3}$$

where (almost as in Lemma 1) for every i, $1 \leq i \leq n$, we have $y_i = \overleftrightarrow{V}_i$ if $v_i \in \mathcal{V}$ and $y_i = \overleftarrow{V}_i\#$ otherwise. Furthermore for every i, $1 \leq i \leq m$, $z_i = \overleftarrow{V}_{j_{2i-1}}\overleftarrow{V}_{j_{2i}}\#$ if the edge $(v_{j_{2i-1}}, v_{j_{2i}})$ is not covered by $\mathcal{V}$. If $v_{j_{2i-1}} \in \mathcal{V}$ then $z_i = \overleftrightarrow{V}_{j_{2i-1}}\overleftarrow{V}_{j_{2i}}$ or $z_i = \overleftarrow{V}_{j_{2i-1}}\overleftrightarrow{V}_{j_{2i}}$ if $v_{j_{2i}} \in \mathcal{V}$. Now, we want to look at the size of G, which is equal to the size of the rules of G and the axiom. The size of the rules is

$$\begin{aligned}\sum_{i=1}^{n} |\overline{v_i}\#| + |\#\overline{v_i}| + \sum_{i=1}^{l} |\#\overline{v_i}\#| + |[\diamond]| &= n(2\lceil \log(n)\rceil + 2) + l(\lceil \log(n)\rceil + 2) + n^3 \\ &= 2n\lceil \log(n)\rceil + 2n + l\lceil \log(n)\rceil + 2l + n^3.\end{aligned}$$

Then, we will look at the size of the axiom;

$$\begin{aligned}|\mathrm{ax}| &= 4n(2\lceil \log(n)\rceil + 3) + (3n - l)(\lceil \log(n)\rceil + 1) + 6t + (m - t)8 + 1 + n^3 \\ &= 8n\lceil \log(n)\rceil + 12n + 3n\lceil \log(n)\rceil + 3n - l\lceil \log(n)\rceil - l + 6t + 8m - 8t + 1 + n^3 \\ &= 11n\lceil \log(n)\rceil + 15n - l\lceil \log(n)\rceil - l + 8m - 2t + 1 + n^3.\end{aligned}$$

Now we have $|G|$ it is

$$\begin{aligned}|G| &= \underbrace{2n\lceil \log(n)\rceil + 2n + l\lceil \log(n)\rceil + 2l + n^3}_{\text{size of rules}} + \underbrace{11n\lceil \log(n)\rceil + 15n - l\lceil \log(n)\rceil - l + 8m - 2t + 1 + n^3}_{|\mathrm{ax}|} \\ &= 13n\lceil \log(n)\rceil + 17n + l + 8m - 2t + 1 + 2n^3.\end{aligned}$$

From the assumptions of the Lemma, we also have that $|G| \leq 13n\lceil \log(n)\rceil + 17n + k + 6m + 1 + 2n^3$. This means that

$$\begin{aligned}13n\lceil \log(n)\rceil + 17n + l + 8m - 2t + 1 + 2n^3 &\leq 13n\lceil \log(n)\rceil + 17n + k + 6m + 1 + 2n^3 \\ l + 8m - 2t &\leq k + 6m \\ l + 2m - 2t &\leq k\end{aligned}$$

Since the total number of edges m is larger or equal to the number of edges covered by $\mathcal{V}$, t, we have that $2m - 2t \geq 0$, which leads to $l \leq k$. It also leads $m - \frac{k-l}{2} \leq t$, meaning that at least $m - \frac{k-l}{2}$ of the edges of $\mathcal{G}$ are covered by $\mathcal{V}$. We denote the number of edges that are still uncovered q. To make a complete vertex cover $\mathcal{V}'$, we must add at most q vertices. Since $q \leq \frac{k-l}{2} \leq k - l$, we get that adding q vertices to $\mathcal{V}$ will influence the size of the new vertex cover the following way

$$|\mathcal{V}'| \leq |\mathcal{V}| + q \leq l + (k - l) = k.$$

This means that we can obtain a vertex cover, $\mathcal{V}'$, of the graph $\mathcal{G}$ that has size at most k. □

Theorem 1. *Even for an alphabet of size 5, the* Smallest 1-Level Grammar Problem *is NP-hard.*

Proof. This follows directly from Lemma 1 and Lemma 6. □

4 Conclusion

We have in this seminar paper shown that the Smallest 1-Level Grammar Problem is NP-hard for even finite alphabets (in this case an alphabet of size 5). The results were first shown in [1] and this seminar paper has recreated and elaborated on the proof from [2].

References

[1] K. Casel, H. Fernau, S. Gaspers, B. Gras, and M. L. Schmid. On the complexity of grammar-based compression over fixed alphabets. *Leibniz International Proceedings in Informatics, Lipics*, 55:122, 2016. ISSN 18688969. doi: 10.4230/LIPIcs.ICALP.2016.122.

[2] K. Casel, H. Fernau, S. Gaspers, B. Gras, and M. L. Schmid. On the complexity of the smallest grammar problem over fixed alphabets. *Theory of Computing Systems*, 65(2):344–409, 2021. ISSN 14330490, 14324350. doi: 10.1007/s00224-020-10013-w.

[3] M. Charikar, E. Lehman, D. Liu, R. Panigrahy, M. Prabhakaran, A. Sahai, and A. Shelat. The smallest grammar problem. *Ieee Transactions on Information Theory*, 51(7):2554–2576, 2005. ISSN 15579654, 00189448. doi: 10.1109/TIT.2005.850116.