THEORY OF COMPUTING, Volume 21 (9), 2025, pp. 1–32
www.theoryofcomputing.org

Complexity of Linear Operators

Alexander S. Kulikov Ivan Mikhailin Andrey Mokhov
Vladimir V. Podolskii

Received December 6, 2020; Revised July 21, 2023; Published November 10, 2025

Abstract. Let A be an n-by-n 0/1-matrix with z zeroes and u ones and let x be an n-dimensional vector of formal variables over a semigroup $(S, \circ)$. How many semigroup operations are required to compute the linear operator Ax?

It is easy to compute Ax using $O(u)$ semigroup operations. The main question studied in this paper is: can Ax be computed using $O(z)$ semigroup operations? For the case when the semigroup is commutative, we give a constructive proof of an $O(z)$ upper bound. This implies that in the commutative settings, the complements of sparse matrices can be processed as efficiently as sparse matrices, though the corresponding algorithms are more involved. This covers the cases of Boolean and tropical semirings that have numerous applications, e. g., in graph theory. On the other hand, we prove that in general this is not possible: for faithful non-commutative semigroups there exists an n-by-n 0/1-matrix with exactly two zeroes in every row (hence $z = 2n$) whose complexity is $\Theta(n\alpha(n))$ where $\alpha(n)$ is the inverse Ackermann function.

As a simple application of the linear-size construction presented, we show how to multiply two $n \times n$ matrices over an arbitrary semiring in $O(n^2)$ time if one of

An extended abstract of this paper appeared in the Proceedings of the 30th International Symposium on Algorithms and Computation (ISAAC), 2019 [17].

ACM Classification: Theory of computation: Streaming, sublinear and near linear time algorithms

AMS Classification: Analysis of algorithms and problem complexity (68Q25)

Key words and phrases: algorithms, linear operators, commutativity, range queries, circuit complexity

 DOI: 10.4086/toc.2025.v021a009

these matrices is a 0/1-matrix with $O(n)$ zeroes (i. e., the complement of a sparse matrix).

1 Introduction

1.1 Problem statement and new results

Let $A \in \{0,1\}^{n\times n}$ be a matrix with z zeroes and u ones, and $x = (x_1, \ldots, x_n)$ be an n-dimensional vector of formal variables over a semigroup $(S, \circ)$. In this paper, we study the complexity of the *linear operator* Ax, i. e., how many semigroup operations are required to compute a vector whose i-th element is

$$\sum_{1\le j\le n \;\wedge\; A_{ij}=1} x_j \tag{1.1}$$

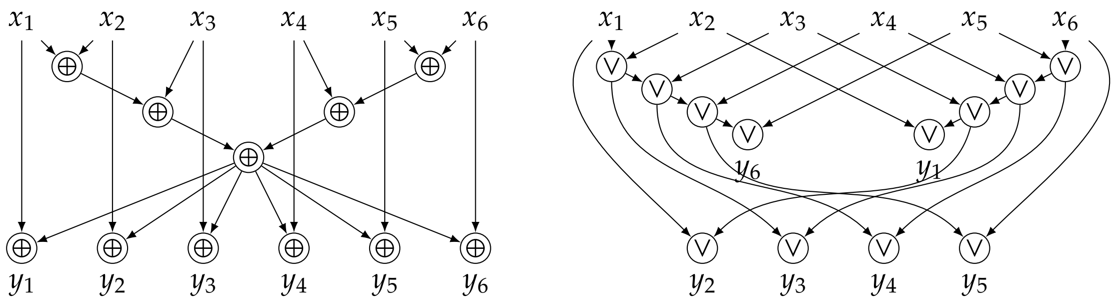
where the summation is over the semigroup operation $\circ$.[1]

To give an example, consider the complement $A \in \{0,1\}^{6\times 6}$ of the identity matrix. In this case, the i-th output y_i (for $i = 1, \ldots, 6$) is equal to the sum of all input variables $x_1, \ldots, x_6$ except for x_i.

x_1	x_2	x_3	x_4	x_5	x_6	
0	1	1	1	1	1	$y_1 = x_2 \circ x_3 \circ x_4 \circ x_5 \circ x_6$
1	0	1	1	1	1	$y_2 = x_1 \circ x_3 \circ x_4 \circ x_5 \circ x_6$
1	1	0	1	1	1	$y_3 = x_1 \circ x_2 \circ x_4 \circ x_5 \circ x_6$
1	1	1	0	1	1	$y_4 = x_1 \circ x_2 \circ x_3 \circ x_5 \circ x_6$
1	1	1	1	0	1	$y_5 = x_1 \circ x_2 \circ x_3 \circ x_4 \circ x_6$
1	1	1	1	1	0	$y_6 = x_1 \circ x_2 \circ x_3 \circ x_4 \circ x_5$

How many operations are required to compute these six sums? The answer depends on the properties of the semigroup S. For example, if $S = (\{0,1\}, \oplus)$, then one can first compute the sum of all input variables a and then let $y_i = a \oplus x_i$. However, this strategy does not work for $S = (\{0,1\}, \vee)$. For this semigroup, one can first compute all prefix sums p_i and suffix sums s_i and then let $y_i = p_{i-1} \vee s_{i+1}$, with appropriate adjustments at the boundaries. See the resulting circuits below.

[1]Note that the result of summation is undefined in case of an all-zero row, because semigroups have no neutral element in general. One can trivially sidestep this technical issue by adding an all-one column $n+1$ to the matrix A, as well as the neutral element x_{n+1} into the vector. Alternatively, we could switch from semigroups to *monoids*, but we choose not to do that, since we have no use for the neutral element and associated laws in the rest of the paper.



In this paper, we are interested in lower and upper bounds involving z and u. Computing all n outputs of Ax directly, i. e., using Definition (1.1) (above), takes $O(u)$ semigroup operations. The main question we study is:

Can Ax be computed using $O(z)$ semigroup operations?

Note that it is easy to achieve $O(z)$ complexity if $\circ$ has an inverse. Indeed, in this case Ax can be computed via subtraction: $Ax = (U - \overline{A})x = Ux - \overline{A}x$, where U is the all-ones matrix whose linear operator can be computed trivially using $O(n)$ semigroup operations, and $\overline{A}$ is the complement of A and therefore has only z ones. Our solution for the above example involving $S = (\{0,1\}, \oplus)$ is obtained in precisely this way, by noticing that $\oplus$ is its own inverse.

1.1.1 Commutative case

Our first main result shows that in the commutative case, the complements of sparse matrices can be processed as efficiently as sparse matrices. Specifically, we prove that if the semigroup is commutative, Ax can be computed in $O(z)$ semigroup operations; or, more formally, there exists a circuit of size $O(z)$ that uses $x = (x_1, \dots, x_n)$ as an input and computes Ax by only applying the semigroup operation $\circ$ (we provide the formal definition of the computational model in Section 2.3). Moreover, the constructed circuits are *uniform* in the sense that they can be generated by an efficient algorithm. Hence, our circuits correspond to an elementary algorithm[2] that uses no tricks like examining the values x_i, i. e., the semigroup operation $\circ$ is applied in a (carefully chosen) order that is independent of the specific input x.

Theorem 1.1. *Let $(S, \circ)$ be a commutative semigroup, and $A \in \{0,1\}^{n \times n}$ be a matrix with $z = \Omega(n)$ zeroes. There exists a circuit of size $O(z)$ that takes a vector $x = (x_1, \dots, x_n)$ of formal variables as an input, uses only the semigroup operation $\circ$ at internal gates, and outputs Ax. Moreover, there exists a randomized algorithm that takes the positions of z zeroes of A as an input and outputs such a circuit in time $O(z)$ with probability at least $1 - O(\log^5 n)/n$. There also exists a deterministic algorithm with running time $O(z + n\log^4 n)$.*

We state the result for square matrices to simplify the presentation. Theorem 1.1 generalizes easily to show that Ax for a matrix $A \in \{0,1\}^{m \times n}$ with $z = \Omega(n)$ zeroes can be computed using

[2] We mostly think of computations over semigroups as circuits. However, whenever we discuss more general algorithms operating with the semigroup, we assume that the algorithm can store semigroup elements and can perform semigroup operation over them in one operation.

$O(m+z)$ semigroup operations. Also, we assume that $z = \Omega(n)$ to be able to state an upper bound $O(z)$ instead of $O(z+n)$. Note that when $z < n$, the matrix A is forced to contain all-one rows that can be computed trivially.

The following corollary generalizes Theorem 1.1 from vectors to matrices.

Corollary 1.2. *Let S be a semiring. There exists a deterministic algorithm that takes a matrix $A \in \{0,1\}^{n\times n}$ with $z = O(n)$ zeroes and a matrix $B \in S^{n\times n}$ and computes the product AB in time $O(n^2)$.*

1.1.2 Non-commutative case

As our second main result, we show that *commutativity is essential*: for any faithful non-commutative semigroup S (the notion of faithful non-commutative semigroup is made formal later in the text in Definition 4.5), the minimum number of semigroup operations required to compute Ax for a matrix $A \in \{0,1\}^{n\times n}$ with $z = O(n)$ zeroes is $\Theta(n\alpha(n,n))$, where $\alpha(n,n)$ is the inverse Ackermann function [19]. For brevity further on we use the notation $\alpha(n) = \alpha(n,n)$.

Theorem 1.3. *There exists a matrix $A \in \{0,1\}^{n\times n}$ with exactly two zeroes in every row such that for any faithful non-commutative semigroup $(S,\circ)$ the minimum number of semigroup operations required to compute Ax is $\Omega(n\alpha(n))$. This lower bound is tight: Ax is computable using $O(n\alpha(n))$ semigroup operations for any $(S,\circ)$ and $A \in \{0,1\}^{n\times n}$.*

The upper bound $O(n\alpha(n))$ follows directly from Yao's result [23]. Hence, our main contribution in the non-commutative case is the lower bound $\Omega(n\alpha(n))$, which we derive from the result of Chazelle and Rosenberg [6].

1.2 Motivation

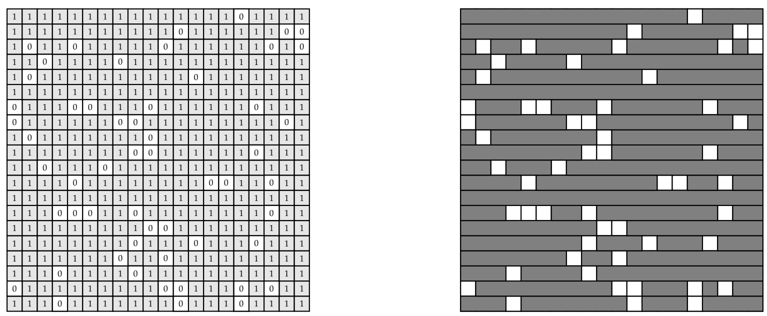
The complexity of linear operators is interesting for many reasons, some of which are listed below.

1.2.1 Range queries

In the *range query* problem, given a vector $x = (x_1, \dots, x_n)$ over a semigroup $(S,\circ)$ and multiple queries of the form (l,r), one is required to output the result $x_l \circ x_{l+1} \circ \cdots \circ x_r$ for each query. It is a classical problem in data structures and algorithms with applications in many fields.

Yao [23] showed that, for any semigroup, it is possible to preprocess the input vector in time $O(n)$ so that any range query can be answered in time $O(\alpha(n))$, where $\alpha(n)$ is the inverse Ackermann function. Yao also proved a matching lower bound. Later, Alon and Schieber [1] studied a more specific question: what is the minimum number of semigroup operations needed at the preprocessing stage for being able to then answer any query in at most k steps? They proved matching lower and upper bounds for every k. As a special case, they showed how to preprocess the input sequence in time $O(n\log n)$ so that one can answer any subsequent query by applying at most one semigroup operation. Chazelle and Rosenberg [6] studied the commutative version of the problem. They proved the $\Omega(n\alpha(n))$ lower bound on the number of commutative semigroup operations required to answer all queries.

The linear operator problem is a natural generalization of the range query problem: each row of the matrix A defines a subset of the elements of x that needs to be summed up and this subset is not required to be a contiguous range. The algorithms (Theorem 1.1) and hardness results (Theorem 1.3) for the linear operator problem presented in this paper are very much inspired by the above-mentioned classic results for the range query problem. The connection to range queries is straightforward: zeros in a row split this row into a collection of ranges in a natural way.



We review applications as well as a rich variety of algorithmic techniques for the range query problem in Sections A.1 and A.2.

1.2.2 Graph algorithms

In this paper, by "graphs" we mean "simple graphs," i. e., undirected graphs without loops and parallel edges.

Many graph problems can be reduced to matrix multiplication. Two classic examples are: (i) the all-pairs shortest path problem (APSP) is reducible to min-plus matrix multiplication [9], and (ii) the number of triangles in a graph can be found by computing the third power (over the integers) of its adjacency matrix [13, 21]. It is natural to ask what happens if a graph or its complement has $O(n)$ edges. (As usual, by n we denote the number of nodes.) In many cases, an efficient algorithm for sparse graphs ($O(n)$ edges) is straightforward whereas an algorithm with the same efficiency for the complements of sparse graphs is not. For example, it is easy to solve APSP and triangle counting on sparse graphs in time $O(n^2)$, but achieving the same time complexity for the complements of sparse graphs is more complicated. Theorem 1.1 and Corollary 1.2 give a black-box way to solve these two problems on the complements of sparse graphs in time $O(n^2)$.

1.2.3 Matrix multiplication over semirings

Fast matrix multiplication methods rely essentially on the ring structure of the underlying set of elements. The first such algorithm was given by Strassen, the current record upper bound is $O(n^{2.373})$ [20, 10]. The removal of the inverse operation often drastically increases the complexity of algorithmic problems over algebraic structures, and even the complexity of standard computational tasks is not well understood over tropical and Boolean semirings (see, e. g., [22, 12]). For various important semirings, we still do not know an $n^{3-\varepsilon}$ (for a constant $\varepsilon > 0$) upper bound for matrix multiplication, e. g., the strongest known upper bound for min-plus matrix multiplication is $n^3/\exp(\sqrt{\log n})$ [22].

The interest in computations over such algebraic structures has recently grown substantially throughout the Computer Science community with the cases of Boolean and tropical semirings being of main interest (see, e. g., [15, 22, 5]). From this perspective, the computational complexity over sparse 0/1-matrices and their complements is one of the most basic questions. Theorem 1.1 and Corollary 1.2 therefore characterise natural special cases when efficient computations are possible.

1.2.4 Functional programming

The *diagonal* $\Delta(V)$ of the set V is the set $\{(x, x) \mid x \in V\}$. By *digraphs* (directed graphs) we mean pairs $G = (V, E)$ of sets where $E \subseteq V \times V$. So, self-loops, i. e., edges in $\Delta(V)$, are permitted. If we wish to exclude such edges, we speak of *loop-free* digraphs. The *complement* of G is $(V, V \times V \setminus E)$. The *loop-free complement* of a loop-free digraph G is $(V, V \times V \setminus \Delta(V) \setminus E)$. Graphs can be viewed as loop-free digraphs where E is a symmetric relation on V. By the *complement of a graph* we mean its loop-free complement.

One of the algebraic data structures developed in the functional programming community for the representation and manipulation of digraphs $G = (V, E)$ is based on the following operations:

- $\langle v \rangle$ (the digraph with the single vertex v and no edges)
- union: $G_1 \cup G_2 = (V_1 \cup V_2, E_1 \cup E_2)$
- join: $G_1 * G_2 = (V_1 \cup V_2, E_1 \cup E_2 \cup (V_1 \times V_2))$, including the resulting self-loops, if any.

Certain classes of digraphs can be generated by these operations in linear (in the number n of vertices) time and memory. These obviously include all sparse digraphs (digraphs with $O(n)$ edges) but they also include digraphs of arbitrary density. For instance, $\langle 1 \rangle * \cdots * \langle n \rangle$ is the transitively oriented complete graph (transitive tournament) which has density 1/2.

It was not known whether the complements of sparse graphs admit such a concise representation. In fact, this specific question motivated our research. Our result gives a constructive positive answer: Theorem 1.1 yields an efficient algorithm for deriving a linear-size algebraic graph representation for the complements of sparse digraphs. It easily follows that the same result is true for graphs and for loop-free digraphs.

1.2.5 Circuit complexity

Computing linear operators over the Boolean semiring $(\{0, 1\}, \vee)$ is a well-studied problem in circuit complexity. The corresponding computational model is known as *rectifier networks*. An overview of known lower and upper bounds for such circuits is given by Jukna [14, Section 13.6]. Theorem 1.1 states that linear operators on the complements of sparse matrices have linear rectifier network complexity.

1.3 Organization and earlier publication

Background definitions are introduced in Section 2. The main results are presented in Section 3 (the commutative case) and Section 4 (the non-commutative case). This paper extends an earlier conference publication [17] by providing complete proofs of all claimed results in Sections 3 and 4.

2 Background

2.1 Semigroups and semirings

A *semigroup* $(S, \circ)$ is an algebraic structure, where the set S is closed under the operation $\circ$, i. e., $\circ : S \times S \to S$, and *associative*, i. e., $x \circ (y \circ z) = (x \circ y) \circ z$ for all x, y, and z in S. *Commutative* (or *abelian*) semigroups introduce one extra requirement: $x \circ y = y \circ x$ for all x and y in S.

A commutative semigroup $(S, \circ)$ can often be extended to a *semiring* $(S, \circ, \bullet)$ by introducing another associative (but not necessarily commutative) operation $\bullet$ that *distributes* over $\circ$, that is

$$x \bullet (y \circ z) = (x \bullet y) \circ (x \bullet z) \qquad\qquad (x \circ y) \bullet z = (x \bullet z) \circ (y \bullet z)\,.$$

hold for all x, y, and z in S. Furthermore, *zero* $0 \in S$ and *one* $1 \in S$ are the *additive* and *multiplicative identities* of the two operators, and zero is *annihilating*:

$$0 \circ x = x \circ 0 = x \qquad 1 \bullet x = x \bullet 1 = x \qquad 0 \bullet x = x \bullet 0 = 0\,.$$

Since $\circ$ and $\bullet$ behave similarly to numeric addition and multiplication, it is common to give $\bullet$ a higher precedence to avoid unnecessary parentheses, and even omit $\bullet$ from formulas altogether, replacing it by juxtaposition. This gives a terser and more convenient notation, for example, the left distributivity law becomes: $x(y \circ z) = xy \circ xz$. We will use this notation, insofar as it does not lead to ambiguity.

2.2 Range query problem and linear operator problem

In the *range query problem*, one is given a sequence $x_1, x_2, \ldots, x_n$ of elements of a fixed semigroup $(S, \circ)$. Then, a *range query* is specified by a pair (l, r) of indices such that $1 \leq l \leq r \leq n$. The answer to such a query is the result of applying the semigroup operation to the corresponding range, i. e., $x_l \circ x_{l+1} \circ \cdots \circ x_r$. The range query problem is then to simply answer all given range queries. There are two regimes: online and offline. In the *online regime*, one is given a sequence of *values* $x_1 = v_1, x_2 = v_2, \ldots, x_n = v_n$ and is asked to preprocess it so that one can efficiently answer any subsequent query. By "efficiently" one usually means in time independent of the length of the range (i. e., $r - l + 1$, the time of a naive algorithm), say, in time $O(\log n)$ or $O(1)$.

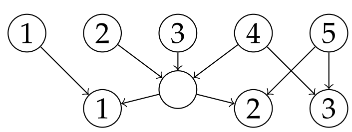
In this paper, we focus on the *offline* version, where one is given a sequence together with all the queries, and are interested in the minimum number of semigroup operations needed to answer all the queries. Moreover, we study a more general problem: we assume that $x_1, \dots, x_n$ are formal variables rather than actual semigroup values. That is, we study the *circuit size* of the corresponding computational problem.

The *linear operator* problem generalizes the range queries problem: now, instead of contiguous ranges one wants to compute sums over arbitrary subsets. These subsets are given as rows of a 0/1-matrix A.

2.3 Circuits

We consider circuits whose input consists of n formal variables $\{x_1, \dots, x_n\}$. We are interested in the minimum number of semigroup operations needed to compute all given words $\{w_1, \dots, w_m\}$ (e. g., for the range query problem, each word has a form $x_l \circ x_{l+1} \circ \dots \circ x_r$). We use the following natural *circuit* model. A circuit computing all these queries is a directed acyclic graph. There are exactly n nodes of zero in-degree. They are labelled with $\{1, \dots, n\}$ and are called *input gates*. All other nodes have positive in-degree and are called *internal gates*. Finally, some m gates have out-degree 0 and are labelled with $\{1, \dots, m\}$; they are called *output gates*. The *size* of a circuit is its number of edges (also called *wires*). Each gate of a circuit computes a word defined in a natural way: input gates compute just $\{x_1, \dots, x_n\}$; any other gate of in-degree r computes a word $f_1 \circ f_2 \circ \dots \circ f_r$ where $\{f_1, \dots, f_r\}$ are words computed at its predecessors (therefore, we assume that there is an underlying order on the incoming wires for each gate). We say that the circuit computes the words $\{w_1, \dots, w_m\}$ if the words computed at the output gates are equivalent to $\{w_1, \dots, w_m\}$ over the semigroup under consideration.

For example, the circuit below computes range queries $(l_1, r_1) = (1, 4)$, $(l_2, r_2) = (2, 5)$, and $(l_3, r_3) = (4, 5)$ over inputs $\{x_1, \dots, x_5\}$ or, equivalently, the linear operator Ax where the matrix A is given below.



$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

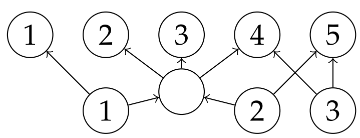
For a 0/1-matrix A, by $C(A)$ we denote the minimum size of a circuit computing the linear operator Ax.

A *binary circuit* is a circuit having no gates of fan-in more than two. It is not difficult to see that any circuit can be converted into a binary circuit of size at most twice the size of the original circuit. For this, one just replaces every gate of fan-in k, for $k > 2$, by a binary tree with $2k - 2$ wires (such a tree contains k leaves hence $k - 1$ inner nodes and $2k - 2$ edges). In the binary circuit the number of gates does not exceed its size (i. e., the number of wires). And the number of gates in a binary circuit is exactly the minimum number of semigroup operations needed to compute the corresponding function.

We call a circuit C computing A *regular* if for every pair (i, j) such that $A_{ij} = 1$, there exists exactly one path from the input j to the output i. A convenient property of regular circuits is the following observation.

Observation 2.1. *Let C be a regular circuit computing a $0/1$-matrix A over a commutative semigroup. Then, by reversing all the wires in C one gets a circuit computing A^T.*

Instead of giving a formal proof, we provide an example of a reversed circuit from the example given above. It is because of this observation that we require circuit outputs to be gates of out-degree zero (so that when reversing all the wires the inputs and the outputs exchange places).



$$A^T = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

3 Commutative case

This section is devoted to the proofs of Theorem 1.1 and Corollary 1.2, which we restate below.

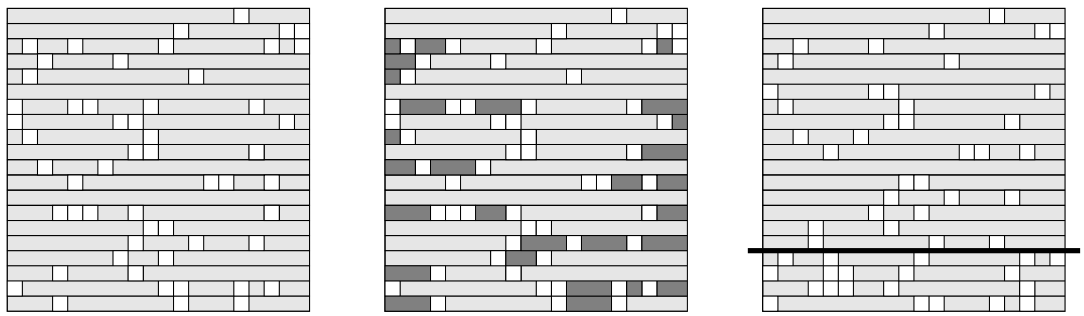
Theorem 3.1 (Theorem 1.1 restated)**.** *Let $(S, \circ)$ be a commutative semigroup, and $A \in \{0,1\}^{n\times n}$ be a matrix with $z = \Omega(n)$ zeroes. There exists a circuit of size $O(z)$ that takes a vector $x = (x_1, \dots, x_n)$ of formal variables as an input, uses only the semigroup operation $\circ$ at internal gates, and outputs Ax. Moreover, there exists a randomized algorithm that takes the positions of z zeroes of A as an input and outputs such a circuit in time $O(z)$ with probability at least $1 - O(\log^5 n)/n$. There also exists a deterministic algorithm with running time $O(z + n\log^4 n)$.*

Corollary 3.2 (Corollary 1.2 restated)**.** *Let S be a semiring. There exists a deterministic algorithm that takes a matrix $A \in \{0,1\}^{n\times n}$ with $z = O(n)$ zeroes and a matrix $B \in S^{n\times n}$ and computes the product AB in time $O(n^2)$.*

3.1 Main ideas of the proof

Consider a matrix $A \in \{0,1\}^{n\times n}$ with $z = \Omega(n)$ zeros (left picture below). Zeros split every row of A into ranges. We construct a circuit of size $O(z)$ that computes all these ranges. Then, by using additional $O(z)$ gates one can compute all outputs of Ax. It is ranges of length at most $\log n$ (middle picture) that make the problem difficult: we prove that one can compute all ranges of length at least $\log n$ using $O(z)$ gates. Using this observation, we proceed as follows. We partition the rows into two parts (right picture): every row in the top part contains at most $\log n$ zeroes, whereas every row in the bottom part contains more than $\log n$ zeroes. The bottom part contains at most $z/\log n$ rows, we transpose it and compute in time $O(z/\log n \cdot \log n) = O(z)$.

For the top part, we employ the commutativity and shuffle the columns. Then, the expected total length of all short ranges is $o(n)$ and one can compute all of them directly.



3.2 Formal proof

We start by proving two simpler statements to show how commutativity is important.

Lemma 3.3. *Let S be a (not necessarily commutative) semigroup and let $A \in \{0,1\}^{n\times n}$ contain at most one zero in every row. Then $C(A) = O(n)$.*

Proof. To compute the linear operator Ax, we first precompute all prefix and suffix sums of $x = (x_1, \dots, x_n)$. Specifically, let $p_i = x_1 \circ x_2 \circ \cdots \circ x_i$. All p_i's can be computed using $(n-1)$ binary gates as follows:

$$p_1 = x_1, p_2 = p_1 \circ x_2, p_3 = p_2 \circ x_3, \dots, p_i = p_{i-1} \circ x_i, \dots, p_n = p_{n-1} \circ x_n \,.$$

Similarly, we compute all suffix sums $s_j = x_j \circ x_{j+1} \cdots \circ x_n$ using $(n-1)$ binary gates. From these prefix and suffix sums all outputs can be computed as follows: if a row of A contains no zeroes, the corresponding output is p_n; otherwise if a row contains a zero at position i, the output is $p_{i-1} \circ s_{i+1}$ (for $i = 1$ and $i = n$, we omit the redundant term). □

In the rest of the section, we assume that the underlying semigroup is commutative. Allowing at most two zeroes per row already leads to a non-trivial problem. Below, we show how to construct a circuit of linear size for this special case (and later on we prove a more general result). It is interesting to compare the following lemma with Theorem 1.3 that states that in the non-commutative setting matrices with two zeroes per row are already non-linear.

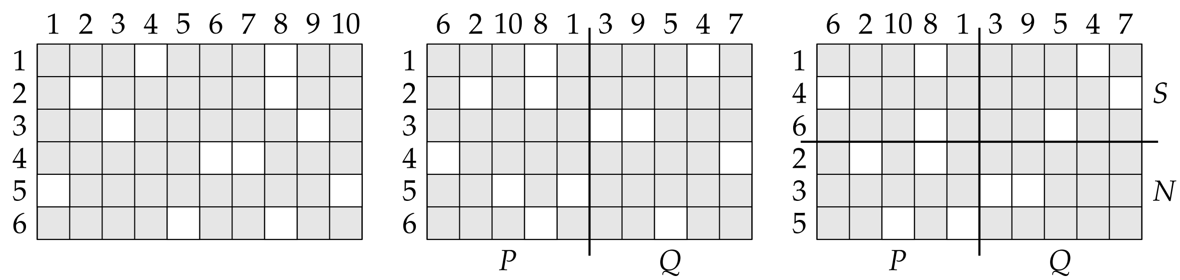
Lemma 3.4. *Let $A \in \{0,1\}^{n\times n}$ contain at most two zeroes in every row. Then $C(A) = O(n)$.*

Proof. Denote by R and C the (sets of) rows and columns of A, respectively. Let $R = R_1 \sqcup R_2$ where every row in R_1 contains at most one zero, whereas every row in R_2 contains exactly two zeros. Clearly, $C(R_1 \times C) = O(n)$ (prefix and suffix sums), hence it remains to prove that $C(R_2 \times C) = O(n)$.

Let $C = C_1 \sqcup C_0$ such that every column of $R_2 \times C_1$ contains at least one zero, whereas $R_2 \times C_0$ is an all-one matrix. It remains to prove that $C(R_2 \times C_1) = O(n)$ gates: after computing

$R_2 \times C_1$, we compute the sum of all variables corresponding to the columns of C_0 (this takes $|C_0| - 1 = O(n)$ gates) and then add this sum to every row of $R_2 \times C_1$ (using $|R_2| = O(n)$ gates). (Working with columns C_1 and C_0 separately is possible due to commutativity.)

To prove that $C(R_2 \times C_1) = O(n)$, we prove that the complexity of a matrix $B = B_R \times B_C \in \{0,1\}^{m\times t}$ containing exactly two zeros in every row and at least one zero in every column (hence, $t \le 2m$) is at most $30m$. We prove this by induction on m. By flipping a coin for every column of B, partition the columns B_C into two parts: $B_C = P \sqcup Q$. We say that $r \in B_r$ is a *split row* if exactly one of two zeros from r lies in P (hence, the other one belongs to Q). For every $r \in B_r$, the probability that r is a split row is $1/2$, hence the expected number of split rows is $|B_R|/2 = m/2$. Take a partition $B_C = P \sqcup Q$ ensuring that the set $S \subseteq B_R$ of split rows has size at least $m/2$ and let $N = B_R \setminus S$ be the set of the non-split rows.



The matrix $S \times B_C$ can be computed by a circuit of size $11m$: each of $S \times P$ and $S \times Q$ has exactly one zero in every column and can be computed using $2t + m$ gates (using prefix and suffix sums); then one more gate suffices for every row; the total size is $(2t + m) + (2t + m) + m \le 11m$.

Thus, it remains to compute the matrix $N \times B_C$. Let $B_C = X \sqcup Y$ where the columns Y do not contain zeros in $N \times B_C$. By induction, the complexity of the matrix $N \times X$ is at most $30|N| \le 30m/2 = 15m$. Then, one computes the sum of all variables from Y (at most $2m$ gates) and adds it to all the rows from N (at most m gates). Thus, the complexity of $N \times B_C$ is at most $18m$.

Overall,

$$C(B) \le C(S \times B_R) + C(N \times B_R) \le 11m + 18m \le 30m\,. \qquad \square$$

Below, we state two auxiliary lemmas that will be used as building blocks in the proof of Theorem 3.1. We prove Lemma 3.6 in Section 3.3.

Lemma 3.5. *There exists a binary regular circuit of size $O(n \log n)$ such that any range can be computed in a single additional binary gate using two gates of the circuit. It can be generated in time $O(n \log n)$.*

Proof. We adopt the divide-and-conquer construction by Alon and Schieber [1]. Split the input range $(1, n)$ into two half-ranges of length $n/2$: $(1, n/2)$ and $(n/2 + 1, n)$. Compute all suffixes of the left half and all prefixes of the right half. Using these precomputed suffixes and prefixes one can answer any query (l, r) such that $l \le n/2 \le r$ in a single additional gate. It remains to be able to answer queries that lie entirely in one of the halves. We do this by constructing recursively circuits for both halves. The resulting recurrence relation $T(n) \le 2T(n/2) + O(n)$ implies that the resulting circuit has size at most $O(n \log n)$. $\square$

Lemma 3.6. *Let $m \le n$ and $A \in \{0,1\}^{m\times n}$ be a matrix with $z = \Omega(n)$ zeroes and at most $\log n$ zeroes in every row. There exists a circuit of size $O(z)$ computing Ax. Moreover, there exists a randomized $O(z)$-time algorithm that takes as input the positions of z zeros and outputs a circuit computing Ax with probability at least $1 - O(\log^5 n)/n$. There also exists a deterministic algorithm with running time $O(n \log^4 n)$.*

Proof of Theorem 3.1. Denote the set of rows and the set of columns of A by R and C, respectively. Let $R_0 \subseteq R$ be all the rows having at least $\log n$ zeroes and $R_1 = R \setminus R_0$. Every row of A can be decomposed into (maximal) contiguous ranges of ones. We call them *ranges of A*. Below, we show that all the ranges of A can be computed by a circuit of size $O(z)$. From these ranges, it takes $O(z)$ additional binary gates to compute all the outputs of Ax.

We compute the matrices $R_0 \times C$ and $R_1 \times C$ separately. The main idea is that $R_0 \times C$ is easy to compute because it has a small number of rows (at most $z/\log n$), while $R_1 \times C$ is easy to compute because it has a small number of zeroes in every row (at most $\log n$).

The matrix $R_1 \times C$ can be computed using Lemma 3.6. To compute $R_0 \times C$, it suffices to compute $C \times R_0$ by a regular circuit, thanks to the Observation 2.1. Let $|R_0| = t$. Clearly, $t \le z/\log n$. Using Lemma 3.5, one can compute all ranges of $C \times R_0$ by a circuit of size

$$O(t\log t + z) = O\left(\frac{z}{\log n}\cdot \log z + z\right) = O(z+n) = O(z)\,,$$

since $z = O(n^2)$.

The algorithm for generating the circuit is just a combination of the algorithms from Lemmas 3.5 and 3.6. ◻

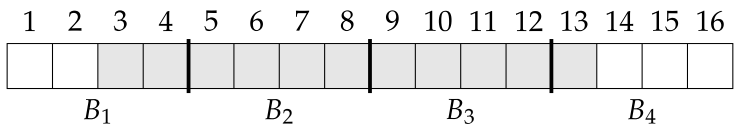
Proof of Corollary 3.2. One deterministically generates a circuit for A of size $O(n)$ in time $O(n\log^4 n) = O(n^2)$ by Theorem 3.1. This circuit can be used to multiply A by any column of B in time $O(n)$. For this, one constructs a topological ordering of the gates of the circuits and computes the values of all gates in this order. Hence, AB can be computed in time $O(n^2)$. ◻

3.3 Deterministic algorithm and the proof of Lemma 3.6

Lemma 3.7. *There exists a binary regular circuit of size $O(n)$ such that any range of length at least $\log n$ can be computed in two additional binary gates from the gates of the circuit. It can be generated by an algorithm in time $O(n)$.*

Proof. We use the block decomposition technique for constructing the required circuit. Partition the input range $(1, n)$ into $n/\log n$ ranges of length $\log n$ and call them blocks. Compute the range corresponding to each block (in total size $O(n)$). Build a circuit from Lemma 3.5 on top of these blocks. The size of this circuit is $O(n)$ since the number of blocks is $n/\log n$. Compute all prefixes and all suffixes of every block. Since the blocks partition the input range $(1, n)$, this also can be done with an $O(n)$ size circuit.

Consider any range of length at least $\log n$. Note that it cannot lie entirely inside the block. Hence, any such range can be decomposed into three components: a suffix of a block, a sequence of whole blocks, and a prefix of a block (where any of the three components may be empty). For example, for $n = 16$, a range $(3, 13)$ is decomposed into a suffix $(3, 4)$ of the first block, a sequence (B_2, B_3) of whole blocks, and a prefix $(13, 13)$ of the last block:



All sequences of blocks can be precomputed by a circuit of size $O(n)$ using the construction from Lemma 3.5 (recall that the number of blocks is $n/\log n$). To combine these three components, one needs two additional binary gates: one to add the suffix, and another to add the prefix. □

Proof of Lemma 3.6. The z zeroes of A break its rows into ranges. Let us call a range *short* is its length is at most $\log n$. Below, we show that it is possible to permute the columns of A so that the total length of all short ranges is at most $o(n)$. Then, all such short ranges can be computed by a circuit of size $o(n) = O(n) = O(z)$. All the remaining ranges can be computed by a circuit of size $O(n)$ using Lemma 3.7.

Randomized algorithm. Permute the columns randomly. A uniform random permutation of n objects can be generated in time $O(n)$ [16, Algorithm P (Shuffling)]. Let us compute the expectation of the total length of short ranges. Let us focus on a single row and a particular cell in it. Denote the number of zeroes in the row by t. What is the probability that the cell belongs to a short segment? There are two cases to consider.

1. The cell is at distance k for $1 \le k \le \log n$ from the border, i. e., it belongs to the first $\log n$ cells or to the last $\log n$ cells (the number of such cells is $2\log n$). Then, this cell belongs to a short range if there is at least one zero in $\log n - k + 1$ cells close to it (on the side opposite to the border). Hence, one zero must belong to the set of $\log n - k + 1$ cells while the remaining $t - 1$ zeroes may be anywhere. The probability is then at most

$$\sum_{1\le k\le \log n} (\log n - k + 1) \cdot \frac{\binom{n}{t-1}}{\binom{n}{t}} \le \log n \cdot \frac{t}{n-t+1} = O\left(\frac{\log^2 n}{n}\right).$$

2. It is not close to the border (the number of such cells is $n - 2\log n$). Then, there must be a zero on both sides of the cell: one at distance $1 \le k < \log n$ on the left and another at distance at most $\log n - k$ on the right. The probability is then at most

$$\sum_{1\le k<\log n} (\log n - k) \cdot \frac{\binom{n}{t-2}}{\binom{n}{t}} \le \log^2 n \cdot \frac{t(t-1)}{(n-t+1)(n-t+2)} = O\left(\frac{\log^4 n}{n^2}\right).$$

Hence, the expected total length of short ranges in one row is

$$O\left(2\log n \cdot \frac{\log^2 n}{n} + (n - 2\log n) \cdot \frac{\log^4 n}{n^2}\right) = O\left(\frac{\log^4 n}{n}\right).$$

Thus, the expected length of short ranges in the whole matrix A is $O(\log^4 n)$. By Markov inequality, the probability that the length of all short ranges is larger than $n/\log n$ is at most $O(\log^5 n/n)$.

Deterministic algorithm. It will prove convenient to assume that A is a $t \times t$ matrix with exactly t zeros with at most $\log t$ zeroes in every row. To do this, we let $t = \max\{n, z\}$ and add a number of all-ones rows and columns if needed. This enlargement of the matrix does not make the computation simpler: additional rows mean additional outputs that can be ignored and additional columns correspond to redundant variables that can be removed (substituted by 0) once the circuit is constructed. Below, we show how to deterministically construct a circuit of size $O(t)$ for A. To do this, we present a greedy algorithm for permuting the columns of A in such a way that the total length of all short segments is $O(\log^4 n)$. This will follow from the fact that all short ranges in the resulting matrix A will lie within the last $O(\log^2 t)$ columns.

We construct the required permutation of columns step by step by a greedy algorithm. After step r, we will have a sequence of the first r columns chosen and we will maintain the following properties:

- For each $i \leq r$, the first i columns contain at least i zeros.
- There are no short ranges within the first r rows (apart from those that can be extended by adding columns on the right).

After $t - \log^2 t$ steps, short ranges will only be possible within the last $\log^2 t + \log t = O(\log^2 t)$ columns. The algorithm itself is presented below.

On the first step, we pick any column that has a zero in it. Suppose we have reached step r. We explain how to add a column on step $r+1$. Consider the last $\log t$ columns in the currently constructed sequence. Consider the set R of rows that have zeros in them. These are exactly the rows that constrain our choice for the next column. There are two cases.

1. There are at most $\log t$ rows in R. Then, for each row in R, there are at most $\log t$ columns that have zeros in this row. In total, there are at most $\log^2 t$ columns that have zeros in some row of R. Denote the set of this columns by F. If there is an unpicked column outside of F that has at least one zero in it, we add this column to our sequence. Clearly, both properties are satisfied and the step is over. Otherwise, all other columns contain only ones, so we add all of them to our sequence, place the columns from F to the end of the sequence, and the whole permutation is constructed.

2. There are more that $\log t$ rows in R. This means that the last $\log t$ columns of the current sequence contain more than $\log t$ zeros. By the first property, the first $r - \log t$ columns contain at least $r - \log t$ zeros. So overall, in the current sequence of r columns there are more than r zeros. Thus, in the remaining $t - r$ columns there are less then $t - r$ zeros and there is a column without zeros. We add this column to the sequence.

To implement this algorithm in time $O(t \log^4 t)$, we store, for each column j of A, a sorted array of rows i such that $A_{ij} = 0$. Since the total number of zeros z is at most $t \log t$, these arrays can be computed in time $O(t\log^2 t)$: if $c_1,\dots,c_t$ are the numbers of zeros in the columns, then sorting the corresponding arrays takes time

$$\sum_{i=1}^{t} c_i \log c_i \le \log(t\log t)\cdot \sum_{i=1}^{t} c_i \le \log(t\log t)\cdot t\log t\,.$$

At every iteration, we need to update the set R. To do this, we need to remove some rows from it (from the column that no longer belongs to the stripe of columns of width $\log t$) and to add the rows of the newly added column. Since the size of $|R|$ is always at most t and the total number of zeros is $z \le t\log t$, the total running time for all such updates is $O(t\log^2 t)$ (if one uses, e. g., a balanced binary search tree for representing R).

If $|R| > \log t$, one just takes an all-one column (all such columns can be stored in a list). If $|R| \le \log t$, we need to find a column outside of the set F. To do this, we just scan the list of the yet unpicked columns. For each column, we first check whether it belongs to the set F. This can be checked in time $O(\log^2 t)$: for every row in $|R|$, one checks whether this row belongs to the sorted array of the considered column using binary search in time $O(\log t)$. Since $|F| \le \log^2 t$, we will find a column outside of F in time $O(\log^4 t)$.

□

4 Non-commutative case

In the previous section, we have shown that for commutative semigroups, co-sparse linear operators can be computed by linear-size circuits. A closer look at the circuit constructions reveals that we use commutativity crucially: it is important that we may reorder the columns of the matrix (we do this in the proof of Lemma 3.6). In this section, we show that this trick is unavoidable: for non-commutative semigroups, it is not possible to construct linear-size circuits for co-sparse linear operators. Specifically, we prove Theorem 1.3 which we restate here.

Theorem 4.1 (Theorem 1.3 restated)**.** *There exists a matrix $A \in \{0,1\}^{n\times n}$ with exactly two zeroes in every row such that for any faithful non-commutative semigroup $(S,\circ)$ the minimum number of semigroup operations required to compute Ax is $\Omega(n\alpha(n))$. This lower bound is tight: Ax is computable using $O(n\alpha(n))$ semigroup operations for any $(S,\circ)$ and $A \in \{0,1\}^{n\times n}$.*

4.1 Faithful semigroups

We consider computations over general semigroups that are not necessarily commutative. In particular, we will establish a lower bound for a large class of semigroups and our lower bound does not hold for commutative semigroups. This requires a formal definition that captures semigroups with rich enough structure and in particular captures the notion that a semigroup is substantially non-commutative.

Previously lower bounds in the circuit model for a large class of semigroups were known for the range query problem [23, 6]. These results were proven for a large class of commutative semigroups called *faithful* (see Definition 4.2). Since we are dealing with the non-commutative case, we need to generalize the notion of faithfulness to non-commutative semigroups.

To provide a formal definition of faithfulness it is convenient to introduce the following notation. Suppose $(S, \circ)$ is a semigroup. Consider variables $x_1, \ldots, x_n$ and consider identities in variables $\{x_1, \ldots, x_n\}$ over $(S, \circ)$. That is, for two words W and W' in the alphabet $\{x_1, \ldots, x_n\}$ we say $W = W'$ iff no matter which elements of the semigroup S we substitute for $\{x_1, \ldots, x_n\}$ we obtain a correct equation over S. Let $X_{S,n}$ be a semigroup with generators $\{x_1, \ldots, x_n\}$ and relations being all identities in variables $\{x_1, \ldots, x_n\}$ over $(S, \circ)$. In other words, $X_{S,n}$ is the quotient of the free semigroup by the congruence relation generated by the given identities. In particular, note that if S is commutative or idempotent then $X_{S,n}$ is also commutative or idempotent, respectively. The semigroup $X_{S,n}$ is studied in algebra under the name of relatively free semigroup of rank n of a variety generated by the semigroup S [18]. We will often omit the subscript n and write simply X_S since the number of generators will be clear from the context. Below we will use the following notation. Let W be a word in the alphabet $\{x_1, \ldots, x_n\}$. Denote by $\mathtt{Var}\,(W)$ the set of letters that are present in W.

We are now ready to introduce, following Yao [23] and Chazelle–Rosenberg [6], the definition of a commutative faithful semigroup.

Definition 4.2 (Yao, Chazelle–Rosenberg)**.** A commutative semigroup $(S, \circ)$ is *faithful commutative* if for any equivalence $W \sim W'$ in X_S we have $\mathtt{Var}\,(W) = \mathtt{Var}\,(W')$.

Note that this definition does not pose any restrictions on the multiplicity of each letter in W and W'. In particular, idempotent semigroups $(\{0, 1\}, \vee)$ and $(\mathbb{Z}, \min)$ are faithful commutative.

We need to study the non-commutative case, and moreover, our results establish the difference between commutative and non-commutative cases. Thus, we need to extend the notion of faithfulness to non-commutative semigroups to capture the whole power of their non-commutativity. At the same time we would like to keep the case of idempotency. We introduce the notion of faithfulness for the non-commutative case inspired by the properties of free idempotent semigroups [11]. To introduce this notion, we need several definitions.

Definition 4.3. The *initial mark* of the non-empty word W W is the letter that is present in W such that its first appearance is farthest to the right. Let U be the prefix of W consisting of the letters preceding the initial mark. That is, U is the maximal prefix of W with a smaller number of generators. We call U the *initial stretch* of W. Analogously we define the *terminal mark* of W and the *terminal* of W.

For example, for $W = abbacabca$, the initial mark is the first letter c, the initial stretch is the prefix $abba$, the terminal mark is the last letter b and the terminal is the suffix ca.

Definition 4.4. We say that a semigroup X with generators $\{x_1, \ldots, x_n\}$ is *strongly non-commutative* if for any words W and W' in the alphabet $\{x_1, \ldots, x_n\}$ the equivalence $W \sim W'$ holds in X only if the initial marks of W and W' are the same, terminal marks are the same, the equivalence $U \sim U'$ holds in X, where U and U' are the initial stretches of W and W', respectively, and the equivalence $V \sim V'$ holds in X, where V and V' are the terminal stretches of W and W', respectively.

In other words, this definition states that the first and the last occurrences of generators in the equivalence separates the parts of the equivalence that cannot be affected by the rest of the generators and must therefore be equivalent themselves. We also note that this definition exactly captures the idempotent case: for a free idempotent semigroup the condition in this definition is "if and only if" [11].

Definition 4.5. A semigroup $(S, \circ)$ is *faithful non-commutative* if X_S is strongly non-commutative.

We note that this notion of faithfulness is relatively general and is true for semigroups $(S, \circ)$ with considerable degree of non-commutativity in their structure. It clearly captures free semigroups with at least two generators. It is also easy to see that the requirements in Definition 4.5 are satisfied for the free idempotent semigroup with n generators (if S is idempotent, then $X_{S,n}$ is also clearly idempotent and no other relations are holding in $X_{S,n}$ since we can substitute generators of S for $x_1, \ldots, x_n$).

When reading through the proof of Theorem 4.1 it is instuctive to keep an example of the free idempotent semigroup in mind. In fact, the very first step of the proof of the lower bound reduces arbitrary semigroup to an idempotent semigroup.

Next we observe some properties of strongly non-commutative semigroups that we need in our constructions.

Lemma 4.6. *Suppose X is strongly non-commutative. Suppose the equivalence $W \sim W'$ holds in X and $|\mathtt{Var}(W)| = |\mathtt{Var}(W')| = k$. Suppose U and U' are minimal (maximal) prefixes of W and W' such that $|\mathtt{Var}(U)| = |\mathtt{Var}(U')| = l \leq k$. Then the equivalence $U \sim U'$ holds in X. The same is true for suffixes.*

Proof. The proof is by induction on the decreasing l. Consider the maximal prefixes first. For $l = k$ and maximal prefixes we just have $U = W$ and $U' = W'$. Suppose the statement is true for some l, and denote the corresponding prefixes by U and U', respectively. Then note that the maximal prefixes with $l - 1$ variables are initial stretches of U and U'. And the statement follows by Definition 4.4.

The proof of the statement for minimal prefixes is completely analogous. Note that on the step of induction the prefixes differ from the previous case by one letter that are initial marks of the corresponding prefixes. So these additional letters are also equal by the Definition 4.4.

The case of suffixes is completely analogous. ◻

The next lemma is a simple corollary of Lemma 4.6.

Lemma 4.7. *Suppose X is strongly non-commutative. Suppose $W \sim W'$ holds in X. Consider a permutation σ_W of the letters of W in the order in which they appear first time in W when we read it from left to right. Consider analogous permutation $\sigma_{W'}$ for W'. Then $\sigma_W = \sigma_{W'}$. The same is true if we read the words from right to left.*

4.2 Proof strategy

We now proceed to the proof of Theorem 4.1. The upper bound follows easily by a naive algorithm: split all rows of A into ranges, compute all ranges by a circuit of size $O(n\alpha(n))$ using Yao's construction [23], then combine ranges into rows of A using $O(n)$ gates.

Thus, we focus on lower bounds. We will view the computation of the circuit as a computation in a strongly non-commutative semigroup $X = X_S$.

We will use the following proof strategy. First we observe that it is enough to prove the lower bound for the case of idempotent strongly non-commutative semigroups X. Indeed, consider an arbitrary semigroup X. Consider a new semigroup X_{id} over the same set of generators that is a factorization of X by idempotency relations $W^2 \sim W$ for all words W in the alphabet $\{x_1, \dots, x_n\}$. We prove the following lemma.

Lemma 4.8. *1. If X is strongly non-commutative, then X_{id} is also strongly non-commutative.*

2. If the co-sparse linear operator problem over X has size s circuit, the co-sparse linear operator problem over X_{id} has size s circuit as well.

As a result a lower bound for the case of X_{id} implies the same lower bound for the case of X. We provide a proof of Lemma 4.8 in Section 4.3.

Hence, from this point we can assume that X is idempotent and strongly non-commutative. Next for idempotent case we show that our problem is equivalent to the commutative version of the range query problem.

For a semigroup X with generators $\{x_1, \dots, x_n\}$ denote by X_{sym} its factorization under commutativity relations $x_i x_j \sim x_j x_i$ for all i, j. Note that if X is idempotent and strongly non-commutative, then X_{sym} is just the semigroup in which $W \sim W'$ iff $\mathtt{Var}(W) = \mathtt{Var}(W')$ (this is free idempotent commutative semigroup).

Theorem 4.9. *For an idempotent strongly non-commutative X and for any $s = \Omega(n)$ we have that the (commutative) range query problem over X_{sym} has size $O(s)$ circuits iff (non-commutative) co-sparse linear operator problem over X has size $O(s)$ circuits.*

For the commutative case it is known that the range query problem is non-linear (Chazelle–Rosenberg [6]).

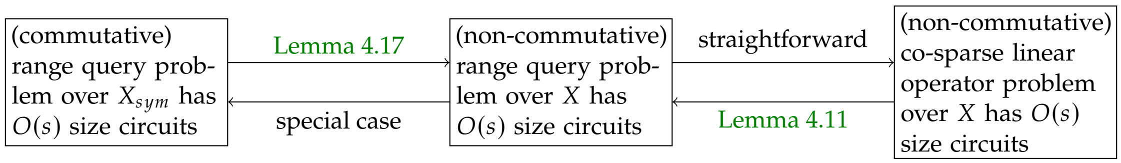
Theorem 4.10 (Chazelle–Rosenberg)**.** *There is a set of n ranges over X_{sym} such that any circuit computing these ranges has size at least $\Omega(n\alpha(n))$.*

Using these results, it is straightforward to finish the proof of Theorem 4.1.

Proof of Theorem 4.1. By Lemma 4.8 it is enough to prove the result for an idempotent strongly non-commutative X. By Theorem 4.9 if non-commutative co-sparse linear operator problem has size s circuit, then the commutative range query problem also does. However, by Theorem 4.10 for the latter problem $s = \Omega(n\alpha(n))$. Moreover, in our construction for the proof of Theorem 4.9 it is enough to consider co-sparse linear operators with exactly two zeroes in every row. From this the lower bound in Theorem 4.1 follows. □

Note that for the proof of Theorem 4.1 only one direction of Theorem 4.9 is needed. However, we think that the equivalence in Theorem 4.9 might be of independent interest, so we provide the proof for both directions.

Thus, it remains to prove Theorem 4.9. We do this by showing the following equivalences for any $s = \Omega(n)$.



In these equivalences, non-commutative problems are considered over an arbitrary strongly non-commutative semigroup and the commutative problem is considered over free idempotent commutative semigroup X_{sym}. Recall that if we factorize any strongly non-commutative idempotent semigroup over commutativity equivalences, we obtain exactly free idempotent commutative semigroup.

Note that two of the reductions on this diagram are trivial. Thus it remains to prove the other two directions.

1. If the (non-commutative) co-sparse linear operator problem over X has size s circuit then the (non-commutative) range query problem over X has size $O(s)$ circuit.

2. If the (commutative) version of the range query problem over X_{sym} has size s circuits then the (non-commutative) version over X also does.

The first of these statements is proved in Sections 4.4 and 4.5. The second statement is proved in Section 4.6.

4.3 From idempotent semigroups to general semigroups

In this section we provide a proof for Lemma 4.8.

First we show that if X is strongly non-commutative, then X_{id} is also strongly non-commutative. Suppose W and W' are words in the alphabet $\{x_1, \dots, x_n\}$ and $W \sim W'$ in X_{id}. This means that there is a sequence $W_0, \dots, W_k$ of words in the same alphabet such that $W = W_0$, $W' = W_k$ and for each i either $W_i \sim W_{i+1}$ in X, or W_{i+1} is obtained from W_i by one application of the idempotency equivalence to some subword of W_i. Clearly, it is enough to check that the conditions of Definition 4.4 are satisfied in X_{id} for each consecutive pair W_i and W_{i+1}.

If $W_i \sim W_{i+1}$ in X, then the conditions of Definition 4.4 follows from the strong non-commutativity of X.

Suppose now that W_{i+1} is obtained from W_i by substituting some subword A by A^2 (the symmetrical case is analyzed in the same way). We will show that initial marks of W_i and W_{i+1} are the same and $U_i \sim U_{i+1}$ in X_{id}, where U_i and U_{i+1} are initial stretches of W_i and W_{i+1} respectively. For the terminals and terminal marks the proof is completely analogous.

Suppose A lies to the left of initial mark in W_i and we substitute A by A^2. Then the initial mark is unaltered and in the initial stretch U_i we also substitute A by A^2. Thus in this case U_{i+1} is obtained from U_i by idempotency relation.

Suppose A contains initial mark of W_i or lies to the right of it. Then after the substitution of A by A^2 the initial mark is still the same and the initial stretch U_i also does not change.

For the second part of the lemma, suppose X is strongly non-commutative and suppose that for X there is a circuit of size at most s computing some co-sparse linear operator. Since X_{id} is a factorization of X any circuit computing co-sparse operator over X also computes the same co-sparse operator over X_{id}. Thus there is a circuit of size at most s computing the same co-sparse operator over X_{id}.

4.4 Reducing co-sparse linear operator to range queries

In this subsection, we prove the following lemma.

Lemma 4.11. *If the (non-commutative) co-sparse linear operator problem over X has size s circuit then the (non-commutative) range query problem over X has size $O(s)$ circuit.*

Intuitively, the lemma holds as the best way to compute rows of a co-sparse matrix is to combine input variables in the natural order. This intuition is formalized in Lemma 4.12 below. Given this, it is easy to reduce co-sparse linear operator problem to the range query problem: we just "pack" each range query into a separate row, i. e., for a query (l, r) we introduce a $0/1$-row having two zeroes in positions $l-1$ and $r+1$ (hence, this row consists of three ranges: $(1, l-1)$, (l, r), $(r+1, n)$). Then, if a circuit computing the corresponding linear operator has a nice property of always using the natural order of variables (guaranteed by Lemma 4.12), one may extract the answer to the query (l, r) from it.

It should be mentioned, at the same time, that the semigroup X might be complicated. In particular, the idempotency is tricky and allows for computations using 'unnatural' order in multiplications. For example, it can be used to simulate commutativity: one can turn xy into yx, by first multiplying xy by y from the left and then multiplying the result by x from the right (obtaining $(y(xy))x = (yx)(yx) = yx$). Using similar ideas, one can place new variables inside of already computed products. To get xyz from xz, one multiplies it by xyz first from the left and then from the right: $(xyz)xz(xyz) = xy(zxzx)yz = xy(zx)yz = xyz$. This is not extremely impressive, since to get xyz we multiply by xyz, but the point is that this is possible in principle.

We proceed to the formal proofs. Let's call a word W in the alphabet $\{x_1, \ldots, x_n\}$ *increasing* if it is a product of variables in the increasing order. A binary circuit is called an *increasing circuit* if each of its gates computes a word equivalent in X to increasing word. Note that if a gate in an increasing circuit is fed by two gates G and H, then the increasing words computed by G and H are matching in a sense that some suffix of G (possibly an empty suffix) is equal to some prefix of H. Otherwise, the result is not equal to a product of variables in the increasing order, due to Lemma 4.7.

Analogously, a binary circuit is called a *range circuit* if each of its gates computes a word that is equivalent to a range.

The proof of Lemma 4.11 follows from the following two lemmas.

Lemma 4.12. *Given a binary circuit computing Ax, one may transform it into an increasing circuit of the same size computing the same function.*

Lemma 4.13. *Given an increasing circuit computing Ax, one may transform it into a range circuit of the same size computing all ranges of A.*

Proof of Lemma 4.11. Given n ranges, pack them into a matrix $A \in \{0,1\}^{n\times n}$ with at most $2n$ zeroes. Take a size-s circuit computing Ax and convert it into a binary circuit. Then, transform it into an increasing circuit using Lemma 4.12. Finally, extract the answers to all the ranges from this circuit using Lemma 4.13. $\square$

Note that the proof of Lemma 4.11 deals with matrices with exactly two zeroes in every row. Thus the lower bound in Theorem 4.1 is true for the same class of matrices.

Next we prove Lemma 4.13 and we prove Lemma 4.12 in the next section.

Proof of Lemma 4.13. Take an increasing circuit C computing Ax and process all its gates in some topological ordering. Each gate G of C computes a (word that is equivalent to an) increasing word. We split this increasing word into ranges and we put into correspondence to G an ordered sequence $G_1, \dots, G_k$ of gates of the new circuit. Each of these gates compute one of the ranges of the word computed by G and $G \sim G_1 \circ \dots \circ G_k$.

Consider a gate G of C and suppose we have already computed all gates of the new circuit corresponding to previous gates of C. G is the product $F \circ H$ of previous gates of C, for which new range gates are already computed. Since C is increasing we have that F and H are matching, that is some suffix (maybe empty) of the increasing word computed in F is equal to some prefix (maybe empty) of the increasing word computed in H and there are no other common variables in these increasing words. It is easy to see that ranges for the sequence corresponding to G are just the ranges for the sequences for F and H with possibly two of them united. If needed, we compute the product of gates of the new circuit corresponding to the united ranges and the sequence of new gates for G is ready.

Thus, to process each gate of C we need at most one operation in the new circuit and the size of the new circuit is at most the size of C.

For output gates of C we have gates in the new circuit that compute exactly ranges of output gates. Thus, in the new circuit all ranges of A are computed. $\square$

4.5 Transforming circuit into an increasing one

In this section we provide a prove for Lemma 4.12.

Consider a binary circuit C computing Ax and its gate G together with a variable x_i it depends on. We say that x_i is *good* in G if there is a path in C from G to an output gate, on which the word is never multiplied from the left by words containing variables greater than or equal to x_i. Note that if x_i and $x_{i'}$ are both contained in G, $i < i'$, and x_i is good in G, then $x_{i'}$ is good in G, too. That is, the set of all good variables in G is closed upwards.

Consider the largest good variable in G (if there is one), denote it by x_k (x_k is actually just the largest variable in G, unless of course there are no good variables in G). Let us focus on the first occurrence of x_k in G.

Claim 4.14. *All first occurrences of other good variables in G must be to the left of the first occurrence of x_k.*

Proof. Suppose that a good variable x_i has the first occurrence to the right of (the first occurrence of) x_k. Consider an output gate H such that there is a path from G to H and along this path there are no multiplications of G from the left by words containing variables greater than x_i. Then we have $H \sim LGR$, where all variables of L are smaller than x_i. Then in H the variable x_i appears before x_k when we read from left to right, but at the same time we have that x_k appears before x_i in LGR. This contradicts Lemma 4.7. □

Now, for a gate G, define two words $\mathtt{MIN}_G$ and $\mathtt{MAX}_G$. Both these words are products of variables in the increasing order: $\mathtt{MIN}_G$ is the product of good variables of G in the increasing order, $\mathtt{MAX}_G$ is the product (in the increasing order) of all variables that has first occurrences before (the first occurrence of) x_k. Note that $\mathtt{MIN}_G$ is a suffix of $\mathtt{MAX}_G$. If there are no good variables in G we just let $\mathtt{MIN}_g = \mathtt{MAX}_g = \lambda$ (the empty word). For the word W that has the form of the product of variables in the increasing order, we call x_j a *gap variable* if it is not contained in W while W contains variables x_i and x_k with $i < j < k$.

Below we show how for a given circuit C to construct an increasing circuit C' that for each gate G of C computes some intermediate product P_G between $\mathtt{MIN}_G$ and $\mathtt{MAX}_G$: $\mathtt{MIN}_g$ is a suffix of P_G and P_G is a suffix of $\mathtt{MAX}_g$. The size of C' is at most the size of C. For an output gate G, $\mathtt{MIN}_g = \mathtt{MAX}_g = g$ hence the circuit C' computes the correct outputs.

To construct C', we process the gates of C in a topological ordering. If G is an input gate, everything is straightforward: in this case $\mathtt{MAX}_G = \mathtt{MIN}_G$ is either λ or x_j. Assume now that G is an internal gate with predecessors F and H. Consider the set of good variables in G. If there are none, we let $P_G = \lambda$. If all first occurrences of good variables of G are lying in one of the predecessors (F and H), then they are good in the corresponding input gate. We then set P_G to P_F or P_H.

The only remaining case is that some good variables have their first occurrence in F while some others have their first occurrence in H. Then the largest variable x_k of G has the first occurrence in H and all variables of F are smaller than x_k.

Claim 4.15. *There are no gap variables for $\mathtt{MAX}_H$ in F.*

Proof. Suppose that some variable x_i in F is a gap variable for $\mathtt{MAX}_H$. Consider an output U such that there is a path from G to U and along this path there are no multiplications of G from the left by words containing variables greater than x_k. Then we have $U \sim LGR$ where all variables of L are smaller than x_k. Consider the prefix P of U preceding the variable x_k and the prefix Q of LG preceding the variable x_k. Then by Lemma 4.6 we have $P \sim Q$. Let us now read P and Q from right to left (note that we switch the order here, previously we read the words from left to right). By Lemma 4.7 the variables in P and Q should appear in the same order. But this is not true (the variable in P are in the decreasing order and in Q the variable x_i is not on its place), a contradiction. □

Claim 4.16. *There are no gap variables for $\mathtt{MAX}_F$ in H.*

Proof. Suppose that a variable x_i in H is a gap variable for $\mathtt{MAX}_F$. Consider an output U such that there is a path from G to U and along this path there are no multiplications of G from the left by words containing variables greater than x_l, the largest variable of F. Then we have $U \sim LGR$, where all variables of L are smaller than x_l. Consider the prefix P of U preceding x_l and the prefix Q of LG preceding x_l. Then by Lemma 4.6 we have $P \sim Q$. But then the variables of P and Q appear in the same order if we read the words from right to left. But this is not true (the variables in P are in the decreasing order and in Q the variable x_i is not on its place), a contradiction. □

We are now ready to complete the proof of Lemma 4.12. Consider P_F and P_H. By Claims 4.15 and 4.16, we know that they are ranges in the same sequence of variables $\mathtt{Var}(P_F) \cup \mathtt{Var}(P_H)$. We know that the largest variables of P_H is greater than all variables of P_F. Then either P_F is contained in P_H, and then we can let $P_G = P_H$ (it contains all good variables of G), or we have $P_F = PQ$ and $P_H = QR$ for some words P, Q, R. In this case we let $P_G = P_F \circ P_H = PQQR = PQR$. Clearly, $\mathtt{MIN}_G$ is the suffix of P_G and P_G itself is the suffix of $\mathtt{MAX}_G$.

4.6 Reducing non-commutative range queries to commutative range queries

In this subsection, we prove the following lemma.

Lemma 4.17. *If the (commutative) version of the range query problem over X_{sym} has size s circuits then the (non-commutative) version over X also does.*

Proof. We will show that any computation of ranges over X_{sym} can be reconstructed without increase in the number of gates in such a way that each gate computes a range (recall, that we call this a range circuit). It is easy to see that then this circuit can be reconstructed as a circuit over X each gate of which computes the same range with the variables in the increasing order. Indeed, we need to make sure that each gate computes a range in such a way that all variables are in the increasing order and this is easy to do by induction. Each gate computes a product of two ranges a and b. If one of them is contained in the other, we simplify the circuit, since the gate just computes the same range as one of its inputs (due to idempotency and commutativity). It is impossible that a and b are non-intersecting and have a gap between them, since then our gate does not compute a range (in a range circuit). So, if a and b are non-intersecting, then they are consecutive and we just need to multiply them in the right order. If the ranges are intersecting, we just multiply them in the right order and apply idempotency.

Thus it remains to show that each circuit for range query problem over X_{sym} can be reconstructed into a range circuit. For this we will need some notation.

Suppose we have some circuit C. For each gate G denote by $\mathtt{left}(G)$ the smallest index of the variable in G (the leftmost variable). Analogously denote by $\mathtt{right}(G)$ the largest index of the variable in G. Denote by $\mathtt{gap}(G)$ the smallest i such that x_i is not in G, but there are some j, k such that $j < i < k$ and x_j and x_k are in G (the smallest index of the variable that is in the gap in G). Next, fix some topological ordering of gates in C (the ordering should be proper, that is inputs to any gate should have smaller numbers). Denote by $\mathtt{num}(G)$ the number of a gate in this ordering. Finally, by $\mathtt{out}(G)$ denote the out-degree of G.

For each gate that computes a non-range consider the tuple

$$\mathtt{tup}(G) = (\mathtt{left}(G), \mathtt{gap}(G), \mathtt{num}(G), -\mathtt{out}(G)) .$$

For the circuit C consider $\mathtt{tup}(C) = \min_G \mathtt{tup}(G)$, where the minimum is considered in the lexicographic order and is taken over all non-range gates. If there are no non-range gates we let $\mathtt{tup}(C) = \infty$. This is our semi-invariant, we will show that if we have a circuit that is not a range circuit, we can reconstruct it to increase its tup (in the lexicographic order) without increasing its size. Since tup ranges over a finite set, we can reconstruct the circuit repeatedly and end up with a range circuit.

Now we are ready to describe a reconstruction of a circuit. Consider a circuit C that is not a range circuit. And consider a gate G such that $\mathtt{tup}(G) = \mathtt{tup}(C)$ (it is clearly unique). Denote by A and B two inputs of G (see Figure 1). Let $i = \mathtt{left}(G)$ and $j = \mathtt{gap}(G)$, that is x_i is the variable with the smallest index in G and x_j is the first gap variable of G (it is not contained in G).

The variable x_i is contained in at least one of A and B. Consider the gate among A and B that contains x_i. This gate cannot have x_j or earlier variable as a gap variable: it would contradict minimality of G (by the second or the third coordinate of tup). Thus this gate is a range $[x_i, x_{j'})$ for some $j' \leq j$ (by this we denote the product of variables from x_i to $x_{j'}$ excluding $x_{j'}$). In particular, only one of A and B contains x_i: otherwise they are both ranges and x_j is not a gap variable for G.

From now on we assume that A contains x_i, that is $A = [x_i, x_{j'})$.

Now we consider all gates $H_1, \ldots, H_k$ that have edges leading from G. Denote by $F_1, \ldots, F_k$ their other inputs. If k is equal to 0, we can remove G and reduce the circuit. Now we consider cases.

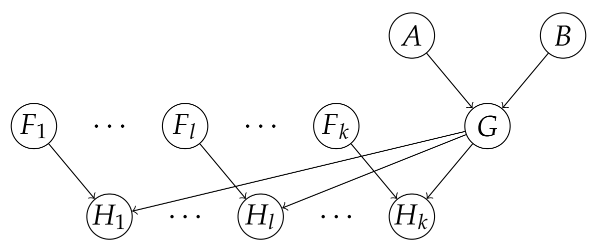


Figure 1: Before reconstruction

Case 1. Suppose that there is l such that $\mathtt{left}(F_l) \leq \mathtt{left}(G)$. If $\mathtt{left}(F_l) < \mathtt{left}(G)$, then F_l must contain all variables $x_i, \ldots, x_j$, since otherwise either F_l or H_l will have smaller tup then G. Thus F_l contains A. Then, we can restructure the circuit by feeding B to H_l instead of G. This does not change the value of the gate computed by H_l and reduces $\mathtt{out}(G)$. Thus $\mathtt{tup}(C)$ increases and we are done.

If $\mathtt{left}(F_l) = \mathtt{left}(G)$, then F_l still cannot have gap variables among $x_i, \ldots, x_{j-1}$ as it would contradict the minimality of G. Thus, F_l is either a range, or it is not a range, but contain all variables $x_i, \ldots, x_{j-1}$. In the latter case again F_l contains A. In the former case F_l either contains A, or is contain in G. If F_l contains A, we can again simplify the circuit as above. If F_l is contained in G, we have $G = H_l$, so we can remove H_l from the circuit and reduce the size of the circuit.

Case 2. Suppose that for all l we have $\mathtt{left}(F_l) > \mathtt{left}(G)$. Consider l such that F_l has the minimal $\mathtt{right}(F_l)$ (if there are several such l pick among them the one with the minimal

$\mathtt{num}(F_l)$). For convenience of notation let $l = k$. Now we restructure the circuit in the following way (see Figure 2). We feed F_k to G instead of A. We feed A to H_k instead of F_k. We feed H_k to all other H_p's instead of G.

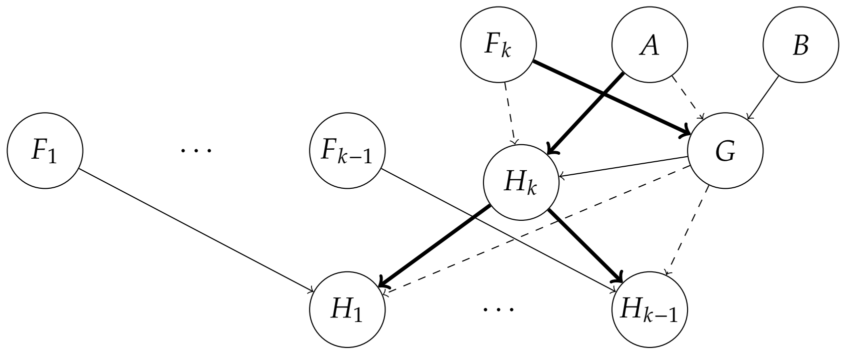


Figure 2: Case 2 reconstruction

Observe that all these reconstructions are valid, that is, they do not create directed cycles in the circuit. To verify this we need to check that there are no cycles using new edges. Indeed, there cannot be a cycle going through one of the edges (H_k, H_p) since this would mean that there was a directed path from H_p to one of the vertices F_k, A and G on the original circuit. Such a path to A or G would mean a cycle in the original circuit. Such a path to F_k violates the minimality property of F_k (minimal $\mathtt{right}(F_k)$). Next, there cannot be a cycle going through both edges (F_k, G) and (A, H_k), since substituting these edges by (F_k, H_k) and (A, G) we obtain one or two cycles in the original circuit. Next, there cannot be a cycle going through the edge (A, H_k) only, since H_k is reachable from A in the original circuit and this would mean a cycle in the original circuit. Finally, there cannot be a cycle going only through the edge (F_k, G) since this would mean a directed path from G to F_k in the original circuit and this contradicts $\mathtt{left}(F_k) > \mathtt{left}(G)$.

Note that our reconstruction might require reordering of the circuit gates, since we create edges between previously incomparable H-gates and between F_k and G. But the reordering affect only the gates with $\mathtt{num}$ greater than $\mathtt{num}(G)$ and may only reduce $\mathtt{num}(F_k)$ to be smaller than $\mathtt{num}(G)$. But this can only increase $\mathtt{tup}(G)$ and since $\mathtt{left}(F_k) > \mathtt{left}(G)$ this can only increase $\mathtt{tup}(C)$.

Observe, that the circuit still computes the outputs correctly. The changes are in the gates $H_1 \ldots, H_k$ (and also in G, but $H_1, \ldots, H_k$ are all of its outputs). H_k does not change. Other H_p's might have changed, they now additionally include variables of F_k. But note that all of these variables are in between of $\mathtt{left}(H_p)$ and $\mathtt{right}(H_p)$, so they must be presented in the output gates connected to H_p anyway (recall that at the output gates we compute ranges).

Now, observe that $\mathtt{tup}(G)$ has increased (by the first coordinate). There are no new gates with smaller $\mathtt{left}$. Among gates with the minimal $\mathtt{left}$ there are no new gates with smaller $\mathtt{gap}$. Among gates with minimal $(\mathtt{left}, \mathtt{gap})$ all gates have larger $\mathtt{num}$ then G. Thus $\mathtt{tup}(C)$ increased and we are done with the proof of Lemma 4.17, completing the proof of Theorem 4.1. □

5 Open problems

There are several natural problems left open.

1. Design a deterministic $O(z)$ time algorithm for generating a circuit in the commutative case. For this, it suffices to design an $O(n)$ deterministic algorithm for the following problem: given a list of positions of n zeroes of an $n \times n$ 0/1-matrix with at most $\log n$ zeroes in every row, permute its columns so that the total length of all segments of length at most $O(\log n)$ is $O(n/\log n)$.

2. Determine the asymptotic complexity of the linear operator in terms of the number of zeroes in the non-commutative case.

3. After the preliminary version of our paper [17], Stasys Jukna posed a question on how large the gap between the complexity of the operators Ax and $\overline{A}x$ can be over the $(\mathbb{N}, +)$ semiring, where $A \in \{0,1\}^{n\times n}$ and $\overline{A}$ is the bitwise negation of A. Our result rules out the possibility of achieving a super-constant (multiplicative) gap with sparse matrix A.

A Review

A.1 Applications of the range query problem

There are many natural applications of the range query problem for a collection of records in a database: computing the total population of cities that are at most some distance away from a given point, computing an average salary in a given period of time, finding the minimum depth on a given subrectangle on a sea map, etc. Below, we review some of the less straightforward applications where efficient algorithms for the range query problem are usually combined with other algorithmic ideas.

String algorithms and computational biology. It is possible to preprocess a given string in $O(n)$ time (where n is its length) so that then to find the longest common prefix of any two suffixes of the original string in constant time. This is done by first constructing the suffix array and the longest common prefix array of the string and then using an efficient RMQ algorithm.

Computational geometry. Algorithms for the range query problem can be used together with the scanning line technique to solve efficiently various problems like: given a set of segments on a line, compute the number of intersecting pairs of segments; or, given a set of rectangles and a set of points on a plane, compute, for each each rectangle, the number of points it contains.

A.2 Known approaches to range queries

In this subsection, we give a brief overview of a rich variety of known algorithms for the range query problem. We say that an algorithm has type $(f(n), g(n))$ if it spends $f(n)$ time on preprocessing the input sequence, and then answers any query in time $g(n)$.

No preprocessing. A naive algorithm skips the preprocessing stage and answers a query (l, r) directly in time $O(r - l + 1)$. It therefore has type $(O(1), O(n))$.

Full preprocessing. One may precompute the answers to all possible queries to be able to answer any subsequent query immediately. Using dynamic programming, it is possible to precompute the answers to all $\Theta(n^2)$ queries in time $O(n^2)$: for this, it is enough to process the queries in order of increasing length. This gives an $(O(n^2), O(1))$ algorithm.

Fixed length queries (sliding window). In case one is promised that all the queries are going to have the same length m, it is possible to do an $O(n)$ time preprocessing and then to answer any query in time $O(1)$. For this, one partitions the input sequence of size n into n/m blocks of size m. For each block, one computes all its prefixes and suffixes in time $O(m)$. The overall running time is $O(n/m \cdot m) = O(n)$. Then, each query of length m touches at most two consecutive blocks and can be answered by taking a precomputed suffix of the left block and a precomputed prefix of the right block in time $O(1)$. This, in particular, implies that, given a sequence of length n and an integer $1 \leq m \leq n$, one may slide a window of length m through the sequence and to output the answer to all such window queries in time $O(n)$.

Prefix sums. In case the semigroup operation has an *easily computable inverse*, there is an $(O(n), O(1))$ algorithm. We illustrate this for a group $(\mathbb{Z}, +)$. Given $x_1, \ldots, x_n$, we compute $(n + 1)$ prefix sums: $S_0 = 0$, $S_1 = x_1$, $S_2 = x_1 + x_2, \ldots, S_n = x_1 + \cdots + x_n$. This can be done in time $O(n)$ since $S_i = S_{i-1} + x_i$. Then, the answer to any query (l, r) is just $S_r - S_{l-1}$.

Note that the algorithm above solves a *static* version of the problem. For the *dynamic* version, where one is allowed to change the elements of the input sequence, there is a data structure known as Fenwick's tree [7]. It allows to change any element as well as to retrieve any prefix sum in time $O(\log n)$.

Block decomposition. One decomposes the input range $(1, n)$ into n/b blocks of length b and then computes, for each block, all its prefixes and suffixes. This can be done in time $O(n)$. Then, for each query, if it lies entirely in a block, we compute the answer directly (hence, in time at most $O(b)$). If it crosses a number of blocks, we decompose it into a suffix of a block, a number of consecutive blocks, and a prefix of a block. This allows us to answer such long queries in time $O(n/b)$. Setting $b = \sqrt{n}$ to balance both cases, we get a $(O(n), O(\sqrt{n}))$-algorithm.

Sparse table. This data structure works for idempotent semigroups (*bands*) and has the type $(O(n \log n), O(1))$. We illustrate its main idea for the *range minimum query* problem (i. e., for a semigroup $(\mathbb{Z}, \min)$). One precomputes answers to $O(n \log n)$ queries—namely, those whose length is a power of 2. More formally, for all $0 \leq k \leq \log_2 n$ and $1 \leq i \leq n - 2^k + 1$, let $S_{k,i}$ be the answer to a query $(i, i + 2^k - 1)$: $S_{k,i} = x_i \circ x_{i+1} \circ \cdots \circ x_{i+2^k-1}$. Since any range of length 2^k consists of two ranges of length 2^{k-1}, one can compute all $S_{k,i}$'s in time $O(n \log n)$ using dynamic programming. Then, any range (l, r) can be covered by two precomputed ranges: if k is the smallest integer such that $2^k \geq (r - l + 1)/2$, then the answer to this query is $S_{k,l} \circ S_{k,r-2^k+1}$ (idempotency is required since we are covering the range, but not partitioning it). This gives an $(O(n \log n), O(1))$ algorithm.

Hybrid strategy. One may extend the block decomposition approach further and use one efficient data structure on top of blocks and possibly a different data structure for each block. Namely, we decompose the input range into blocks of size b, use a $(p_1(n), q_1(n))$-algorithm on top of blocks and a $(p_2(n), q_2(n))$-algorithm within each block. The resulting algorithm then has type

$$(O(n + p_1(n/b) + (n/b) \cdot p_2(b), O(q_1(n/b) + q_2(b))) .$$

For example, for the range minimum problem, combining the sparse table data structure ($p_1(n) = O(n \log n)$, $q_1(n) = O(1)$) with no preprocessing technique ($p_2(n) = O(1)$, $q_2 = O(n)$) and block size $b = \log n$, gives an $(O(n), O(\log n))$-algorithm. Another example: using sparse table in both cases (with block size $b = \log n$) gives an $(O(n \log \log n), O(1))$ algorithm.

Segment tree. The segment tree data structure is also based on dynamic programming ideas and works for any semigroup. Consider the following complete binary tree with $O(n)$ nodes: the root is labelled by a query $(1, n)$, the two children of each inner node (l, r) are labelled by the left and right halves of the current query (i. e., (l, m) and $(m + 1, r)$ where $m = (l + r)/2$), the leaves are labelled by length one queries. Going from leaves to the root, one can precompute the answers to all the queries in this tree in time $O(n)$. Then, it is possible to show that any query (l, r) can be partitioned into $O(\log n)$ queries that are stored in the tree. This gives an $(O(n), O(\log n))$ algorithm. It should be noted that the segment tree can also be used to solve the dynamic version of the range query problem efficiently: to change the value of one of the elements of the input sequence, one needs to adjust the answers to $O(\log n)$ queries stored in the tree.

Algorithms by Yao and by Alon and Schieber. Yao [23] showed that, for any semigroup, it is possible to preprocess the input sequence in time $O(n)$ so that any range query can be answered in time $O(\alpha(n))$ where $\alpha(n)$ is the inverse Ackermann function and proved a matching lower bound. Later, Alon and Schieber [1] studied a more specific question: what is the minimum number of semigroup operations needed at the preprocessing stage for being able to then answer any query in at most k steps? They proved matching lower and upper bounds for every k. As a special case, they show how to preprocess the input sequence in time $O(n \log n)$ so that any subsequent query can be answered by applying at most one semigroup operation. This algorithm generalizes the sparse table data structure (as it does not require the semigroup to be idempotent) and is particularly easy to describe. It is based on the divide-and-conquer paradigm. Let $m = n/2$. We precompute answers to all queries of the form (i, m) and $(m + 1, j)$, where $1 \leq i \leq m$ and $m + 1 \leq j \leq n$ (i. e., suffixes of the left half and prefixes of the right half). This allows to answer in a single step any query that intersects the middle of the sequence, i. e., queries (l, r) such that $l \leq m \leq r$. All the remaining preprocessing boils down to answering queries that lie entirely in either left or right half. This can be done recursively for the halves. The corresponding recurrence relation $T(n) = 2T(n/2) + O(n)$ implies an upper bound $O(n \log n)$ on preprocessing time (and hence, the number of semigroup operations).

$(O(n), O(1))$**-type algorithms.** There is a sequence of $(O(n), O(1))$-type algorithms designed specifically for the range minimum query problem and a related problem called least common ancestor (LCA) [4, 3, 2, 8]. Here, we briefly sketch the algorithm by Bender and Farach-Colton. Its main idea is to first reduce RMQ to LCA (the least common ancestor problem). One then reduces LCA back to RMQ and notices that the resulting instance of RMQ has a convenient property: the difference between any two consecutive elements is ±1. This property allows to do the following trick: we precompute answers to all relatively short queries (this can be done even without knowing the input sequence because of the ±1 property); we also partition the input sequence into blocks and build a segment tree out of these blocks.

Acknowledgments

We thank Paweł Gawrychowski for pointing us to the paper [6], and Alexey Talambutsa for fruitful discussions on the theory of semigroups. We are also grateful to the reviewers for their thorough reviews that helped us improve the final version of the paper. We are grateful to László Babai for numerous suggestions that greatly improved the readability of the paper.

References

[1] NOGA ALON AND BARUCH SCHIEBER: Optimal preprocessing for answering on-line product queries. Technical Report 71/87, Inst. Computer Science, Tel Aviv University, 1987. [arXiv:2406.06321] 4, 11, 28

[2] MICHAEL A. BENDER AND MARTIN FARACH-COLTON: The LCA problem revisited. In *Proc. Latin American Symp. on Theoretical Informatics (LATIN'00)*, pp. 88–94. Springer, 2000. [doi:10.1007/10719839_9] 28

[3] MICHAEL A. BENDER, MARTIN FARACH-COLTON, GIRIDHAR PEMMASANI, STEVEN SKIENA, AND PAVEL SUMAZIN: Lowest common ancestors in trees and directed acyclic graphs. *J. Algorithms*, 57(2):75–94, 2005. [doi:10.1016/j.jalgor.2005.08.001] 28

[4] OMER BERKMAN AND UZI VISHKIN: Recursive star-tree parallel data structure. *SIAM J. Comput.*, 22(2):221–242, 1993. [doi:10.1137/0222017] 28

[5] PETER BUTKOVIČ: *Max-linear Systems: Theory and Algorithms*. Springer, 2010. [doi:10.1007/978-1-84996-299-5] 6

[6] BERNARD CHAZELLE AND BURTON ROSENBERG: The complexity of computing partial sums off-line. *Internat. J. Comput. Geom. Appl.*, 1(1):33–45, 1991. [doi:10.1142/S0218195991000049] 4, 15, 16, 18, 29

[7] PETER M. FENWICK: A new data structure for cumulative frequency tables. *Softw., Pract. Exper.*, 24(3):327–336, 1994. [doi:10.1002/spe.4380240306] 27

[8] JOHANNES FISCHER AND VOLKER HEUN: Theoretical and practical improvements on the RMQ-problem, with applications to LCA and LCE. In *Proc. Annual Symp. on Combinatorial Pattern Matching (CPM'06)*, pp. 36–48. Springer, 2006. [doi:10.1007/11780441_5] 28

[9] MICHAEL J. FISCHER AND ALBERT R. MEYER: Boolean matrix multiplication and transitive closure. In *Proc. 12th Annual Symp. on Switching and Automata Theory (SWAT'71)*, pp. 129–131. IEEE Comp. Soc., 1971. [doi:10.1109/SWAT.1971.4] 5

[10] François Le Gall: Powers of tensors and fast matrix multiplication. In *Proc. 39th Internat. Symp. Symbolic and Algebraic Computation (ISSAC'14)*, pp. 296–303. ACM Press, 2014. [doi:10.1145/2608628.2608664] 5

[11] James A. Green and David Rees: On semi-groups in which $x^r = x$. *Math. Proc. Cambridge Phil. Soc.*, 48(1):35–40, 1952. [doi:10.1017/S0305004100027341] 16, 17

[12] Dima Grigoriev and Vladimir V. Podolskii: Complexity of tropical and min-plus linear prevarieties. *Comput. Complexity*, 24(1):31–64, 2015. [doi:10.1007/s00037-013-0077-5] 6

[13] Alon Itai and Michael Rodeh: Finding a minimum circuit in a graph. *SIAM J. Comput.*, 7(4):413–423, 1978. [doi:10.1137/0207033] 5

[14] Stasys Jukna: *Boolean Function Complexity - Advances and Frontiers*. Volume 27 of *Algorithms and Combinatorics*. Springer, 2012. [doi:10.1007/978-3-642-24508-4] 7

[15] Stasys Jukna: Tropical complexity, Sidon sets, and dynamic programming. *SIAM J. Discr. Math.*, 30(4):2064–2085, 2016. [doi:10.1137/16M1064738] 6

[16] Donald Ervin Knuth: *The Art of Computer Programming, Volume II: Seminumerical Algorithms, 3rd Edition*. Addison-Wesley, 1997. Available via ACM DL. 13

[17] Alexander S. Kulikov, Ivan Mikhailin, Andrey Mokhov, and Vladimir Podolskii: Complexity of linear operators. In *Proc. Internat. Symp. on Algorithms and Computation (ISAAC'19)*, pp. 17:1–12. Springer, 2019. 1, 7, 26

[18] Hanna Neumann: *Varieties of Groups*. Springer, 1967. [doi:10.1007/978-3-642-88599-0] 16

[19] Robert Endre Tarjan: Efficiency of a good but not linear set union algorithm. *J. ACM*, 22(2):215–225, 1975. [doi:10.1145/321879.321884] 4

[20] Virginia Vassilevska Williams: Multiplying matrices faster than Coppersmith–Winograd. In *Proc. 44th STOC*, pp. 887–898. ACM Press, 2012. [doi:10.1145/2213977.2214056] 5

[21] Virginia Vassilevska Williams and R. Ryan Williams: Subcubic equivalences between path, matrix, and triangle problems. *J. ACM*, 65(5):27:1–38, 2018. Preliminary version in FOCS'10. [doi:10.1145/3186893] 5

[22] R. Ryan Williams: Faster all-pairs shortest paths via circuit complexity. *SIAM J. Comput.*, 47(5):1965–1985, 2018. Preliminary version in STOC'14. [doi:10.1137/15M1024524] 6

[23] Andrew Chi-Chih Yao: Space-time tradeoff for answering range queries (extended abstract). In *Proc. 14th STOC*, pp. 128–136. ACM Press, 1982. [doi:10.1145/800070.802185] 4, 15, 16, 17, 28

AUTHORS

Alexander S. Kulikov
Steklov Mathematical Institute at St. Petersburg
Russian Academy of Sciences
alexander.s.kulikov@gmail.com
https://alexanderskulikov.github.io/

Ivan Mikhailin
Steklov Mathematical Institute at St. Petersburg
Russian Academy of Sciences
ivmihajlin@gmail.com
https://dblp.org/pid/40/11440.html

Andrey Mokhov
Jane Street Singapore
and School of Engineering
Newcastle University, U. K.
andrey.mokhov@ncl.ac.uk

Vladimir V. Podolskii
Tufts University
and Steklov Mathematical Institute
Russian Academy of Sciences
vladimir.podolskii@tufts.edu
https://engineering.tufts.edu/cs/people/faculty/vladimir-podolskii

ABOUT THE AUTHORS

ALEXANDER KULIKOV holds Ph. D. (2009) and Dr. Sci. (2017) degrees from the St. Petersburg Department of the Steklov Mathematical Institute. His Ph. D. advisor was Edward A. Hirsch. Currently, Alexander is a researcher at JetBrains Research and the Head of the Computer Science and Artificial Intelligence B. Sc. program at Neapolis University Pafos. His scientific interests include algorithms, circuit complexity, and Computer Science education. He coauthored three books and sixteen massive open online courses on algorithms and discrete mathematics with over a million enrolled students. In his spare time, he enjoys discussing the circuit complexity of the MOD_3 function with Alexander Golovnev.

IVAN MIKHAILIN is a researcher at JetBrains Research. His research areas are circuit complexity, fine-grained complexity, and algorithmic theory. Ivan did his Ph. D. studies (2014–2019) at the University of California San Diego under the supervision of Russell Impagliazzo.

ANDREY MOKHOV is a software engineer at Jane Street Singapore, and a visiting fellow at Newcastle University, UK. His research interests are in applying abstract mathematics and functional programming to solving large-scale engineering problems. During his Ph. D. studies (2005–2009), Andrey worked on asynchronous circuits and concurrency theory under the supervision of Alex Yakovlev. In 2014, he became interested in functional programming and software build systems, which eventually led him to writing more and more code, and in 2019, he switched from academia to industry, joining the Jane Street's Tools and Compilers team. Andrey is originally from Kyrgyzstan where he helps to run the ACM ICPC regional programming contest.

VLADIMIR PODOLSKII defended his Ph. D. thesis in 2009 at Moscow State University advised by Nikolay Vereshchagin. He defended his Dr. Sci. thesis in 2021 at Steklov Mathematical Institute. His research areas are circuit complexity, its applications to databases augmented with ontologies, min-plus geometry. He is an Associate Professor at Tufts.