Approaches to Lower Bounding Cohn-Umans Matrix Multiplication

Khai Dong

November 27, 2023

Abstract

Bounding matrix multiplication runtime has been an active area of research since 1969 when Strassen showed $\omega \leq 2.81$ [18]. In 2003, Cohn and Umans [8] produced a matrix multiplication framework and conjectured $\omega = 2$. Blasiak et al. [5], Sawin [17], and Blasiak et al. [6] rejected this conjecture, showing that for no fixed parameters we can obtain $\omega = 2$ using Cohn and Umans' method. However, a concrete lower bound greater than 2 has never been given. More works over the year by Alman and Williams [1] and Vassilevska Williams [19] focus on lower bounding matrix multiplication using various methods including Cohn and Umans'. This report aims to give a summary and an approach to lower bounding Cohn and Umans' matrix multiplication.

Contents

1 Introduction

Matrix multiplication is a fundamental operation in linear algebra that lies at the center of computation in engineering, computer graphics, and quantum physics. As such, over the years researchers have sought ways to improve this operation. Improving the speed of matrix multiplications may facilitate computing problems that were previously hard to efficiently compute within our current technological limit.

Matrix multiplication is relatively short to define (see Section 2.2). However, the question of how fast matrix multiplication can be is a huge challenge.

1.1 Prior Works

For brevity, we define $\mathcal{O}(n^{\omega})$ to be the runtime of the best matrix multiplication algorithm for $n \times n$ matrices (see Section 2.2.1 for formal definition). It is widely conjectured that $\omega = 2$.

For years, it is only known that $\omega \leq 3$ by the existence of a naive algorithm. The first improvement is made by Strassen [18], yielding $\omega \leq 2.8074$. This was a huge discovery that fueled the research into producing faster matrix multiplication algorithms and raised the question of determining ω. In 1990, Strassen [18]'s approach is further improved by Coppersmith and Winograd [10], yielding $\omega \leq 2.3755$. However, in the 2020s, progress on upper bounding ω has been slow. In 2023, Vassilevska Williams et al. [20] showed $\omega \leq 2.371552$ which is a ten-thousandth improvement from the previous $\omega \leq 2.371866$ showed by Duan et al. [13] in 2022.

In 2003, Cohn and Umans [8] proposed a different approach using a certain abstract algebra structure, conjecturing $\omega = 2$ with their approaches. Cohn et al. [9] further built upon their previous approaches. This approach is the current focus in Professor Anderson's lab, yielding $\omega \leq 2.66$ Anderson et al. [3] and $\omega \leq 2.505$ Anderson and Le [4]. Despite not producing much progress, the framework has been fully implemented, first partially by Anderson et al. [3], and fully by Dubinsky [14] and Dong [12]. This implementation, however, is inefficient due to the large size of relevant structures which does not allow us to produce any non-trivial upper bounds for ω. This directs our research towards either improving the implementation or obtaining a lowerbound to see if we have hit the limit of this method.

This research follows the latter direction, aiming to produce a lower bound greater than 2 for Cohn & Umans' matrix multiplication. This work builds up on top of the works by Blasiak et al. [5], Sawin [17], and Blasiak et al. [6], which rejected Cohn and Umans' conjecture, showing that for no fixed parameters we can obtain $\omega = 2$ using Cohn & Umans' method. They, however, do not provide a concrete tighter lower bound for Cohn and Uman's method since they do not rule out the existence of a family of Cohn & Umans' matrix multiplication algorithms whose runtime approach 2.

1.2 Current Progress

For the first 10 weeks of this research, I have managed to

- Study the relevant backgrounds of this topic. This can be found in Chapters 14 and 15 of Bürgisser et al. [7].
- Review the basic concepts and the abstract algebra involved in the Cohn-Umans method [8, 9].
- Review the relevant research on the topic of lower bounding matrix multiplication framework in general. Some of which are Blasiak et al. [5], Blasiak et al. [6], Alman and Williams [1], and Vassilevska Williams [19].
- Make some progress in linking the background information and relevant research back to my original goal of lower bounding Cohn-Umans method, which will be described in Section 4.

This project is theoretical and requires an understanding of advanced abstract algebra. Section 2 explains these concepts. Then, Section 3 describes Cohn & Umans' matrix multiplication framework, and Section 5 discusses an extensive future plan for this project.

2 Preliminaries

This section served to provide some basics to the reader. For brevity, all proofs are omitted. The same information can also be found in Cormen et al. [11], Bürgisser et al. [7], and Dummit and Foote [15]. We assume the reader knows all basic math including calculus, sets, functions, and equivalent relations, which are covered in MTH113 (or MTH110&MTH112) and MTH199 (or MTH197) which is a requirement for a Computer Science major/minor at Union College. Note that some of the building blocks have been changed to be suited for the general audience. For precise definitions, please check with the references above.
For this project, we denote $[n] := \{1, 2, \ldots, n\}$ where $n \in \mathbb{N}$.

2.1 Abstract Algebra Prerequisites

This section provides some of the building blocks for us to understand upcoming subsections and Cohn and Umans' matrix multiplication.

2.1.1 Groups, Rings, and Fields

Definition 2.1 *A **group** $(G, *)$ is a non-empty set of elements defined under a binary operation $*$ satisfying the following properties:*

- *closure: for all $a, b \in G$, $a * b \in G$,*
- *associative: for all $a, b, c \in G$, $(a * b) * c = a * (b * c)$,*
- *identity: there exists $e \in G$ s.t., for all $a \in G$, $a * e = e * a = a$,*
- *inverse: for all $a \in G$, there exists $a^{-1} \in G$ s.t., $a * a^{-1} = e$.*

When the operation is clear from context, we only write G to stand for the group.

In the context of this project, we will only consider finite groups. Informally, a group is an abstraction of familiar mathematical objects. One example group is $\mathbb{Z}$, the set of integers under $+$. We can check that $(\mathbb{Z}, +)$ is indeed a group: the sum of two integers is an integer, the operation $+$ is associative, trivially, the identity elements of $(\mathbb{Z}, +)$ is 0, and for any integer z, $z + (-z) = 0$.
However, $(\mathbb{Z}, +)$ is a rather special case. Note that in this definition, there is no mention of elements commuting through the operation. This means in a group $a * b$ is not necessarily equal to $b * a$. If every element commutes, the group is called **abelian**.

Definition 2.2 *A group G where every element commutes i.e., $a * b = b * a$ for all $a, b \in G$ is called an **abelian** group.*

In $(\mathbb{Z}, +)$, only addition and subtraction (adding the inverse) are defined. With integers, we can also multiply. Thus, we defined a **ring**, which allows multiplications alongside addition and subtraction.

Definition 2.3 *A **ring** $(R, +, *)$ is a set of elements endowed with 2 binary operations $+$ and $*$ such that*

- *$(R, +)$ is an abelian group,*
- *$*$ is closed and associative,*
- *$*$ is distributive over $+$. That is*

$$\forall a, b, c \in R, (a + b) * c = a * c + b * c \quad \textit{and} \quad c * (a + b) = c * a + c * b. \tag{1}$$

Similar to groups, when the operations are clear from context, we denote R as the ring.

A ring is an abstract object that is closer to our usual number fields. However, some basic properties have yet to be included:

- Whether multiplication is commutative
- Whether the multiplicative identity ”1” exists at all

- Whether we can divide (multiply with inverses).

Definition 2.4 *A ring R is* ***commutative*** *if $*$ is commutative.*

Definition 2.5 *A ring R is* ***with unity*** *if there exists $1 \in R$ s.t., for all $a \in R$, $1 * a = a * 1 = a$.*

Definition 2.6 *A ring R is a* ***division ring*** *if it is with unity and for all non-zero $a \in R$, there exists a^{-1} s.t., $a * a^{-1} = a^{-1} * a = 1$.*

Definition 2.7 *A* ***field*** *is a commutative division ring.*

A field is where we can add, subtract, divide, and multiply, therefore mirroring the basic operations on basic number fields such as $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$. Then, we define what it means for 2 rings (or fields) to be the ″same″.

Definition 2.8 *Two rings R and S are said to be* ***isomorphic****, denoted $R \cong S$, if there exists an a bijection $\phi : R \to S$ such that for all $a, b \in R$,*

$$\phi(a) +_S \phi(b) = \phi(a +_R b) \quad \textit{and} \quad \phi(a) *_S \phi(b) = \phi(a *_R b) \tag{2}$$

In this case, ϕ is called a ***ring isomorphism****. Since fields are also rings,* ***field isomorphisms*** *are ring isomorphisms. If ϕ is not a bijection, but still satisfies the secondary property in Equation 2, we call ϕ a* ***ring homomorphism****.*

Informally, if 2 rings R and S are isomorphic, we can obtain S by renaming the elements of R using ϕ. This means the 2 rings are essentially the same object. Here are some examples to build intuition for these objects:

- $\mathbb{Z}$ is a ring with unity but is not a division ring since $2 \in \mathbb{Z}$, but $\frac{1}{2} \notin \mathbb{Z}$.
- $2\mathbb{Z} = \{2z : z \in \mathbb{Z}\}$ is a ring without unity since 1 is odd.
- The set of $n \times n$ matrices is a ring with unity where $*$ is matrix multiplication. It is a fact that matrix multiplication does not commute in general and not all matrices have an inverse.
- The set of $n \times n$ invertible matrices is only a division ring since matrix multiplication does not necessarily commute.
- $\mathbb{Z}_p$ (integers modulo p) is a field if p is prime. This is a finite field.

2.1.2 Polynomial Rings, Group Algebra, Field Extensions, and Closed Fields

Here, we explore some ring/field operations that preserve the ring's properties. As from the previous section, these operations mirror what can be done in the usual fields $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$. Indeed, the following operation defined how we construct polynomials from a ring.

Definition 2.9 *Let R be a ring and x be an indeterminate. The* ***polynomial ring*** *$R[x]$ in x over ring R is the set of all formal sums of the form:*

$$a_n x^n + a_{n-1}x^{n-1} + a_{n-1}x^{n-1} + \cdots + a_0$$

or, in other notation, $\sum_{i=0}^{n} a_i x^i$ for some $n \geq 0$, $a_i \in R$ for $i \in [0 \ldots n]$, and $a_n \neq 0$. For this polynomial, n is called the ***degree*** *of the polynomial.*

Theorem 2.1 *Let R be a ring and x be an indeterminate. The polynomial ring $R[x]$ is a ring.*

Using $R[x]$ as our ring and y as our indeterminate, we can define the polynomial ring $R[x][y]$, also denoted $R[x,y]$, similarly and get another polynomial ring.

Corollary 2.1.1 *Given a ring R and a finite list of indeterminates x_1, x_2, ..., x_n, $R[x_1, x_2, \ldots, x_n]$ is a ring.*

Then, we explore a different property of basic fields like $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$. Firstly, we note that $\mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$. We say $\mathbb{R}$ is a **field extension** of $\mathbb{Q}$, and $\mathbb{C}$ is a **field extension** of $\mathbb{R}$.

Definition 2.10 *Let $\mathbb{K}$ and $\mathbb{L}$ be fields. If $\mathbb{L} \subseteq \mathbb{K}$ s.t, operations of $\mathbb{K}$ are those of Ł, $\mathbb{K}$ is a* ***field extension*** *of $\mathbb{L}$ or, equivalently, $\mathbb{L}$ is a* ***subfield*** *of $\mathbb{K}$.*

Consider the polynomial $f(x) = x^2 - 2$ in $\mathbb{Q}[x]$ with roots $\pm\sqrt{2} \notin \mathbb{Q}$. In this case, we say $\pm\sqrt{2}$ are **algebraic** over $\mathbb{Q}$ as $\pm\sqrt{2} \notin \mathbb{Q}$, but are roots of a polynomial in $\mathbb{Q}[x]$. Or more generally, we have the following definition.

Definition 2.11 *Let $\mathbb{K}$ and $\mathbb{L}$ be fields. Suppose $\mathbb{K}$ is a field extension of $\mathbb{L}$. Let $\alpha \in \mathbb{K}$. If α is a root of a polynomial on $\mathbb{L}[x]$, we say α is* ***algebraic*** *over $\mathbb{L}$.*

We can then apply the same techniques to fields (which are rings by definition).

Definition 2.12 *Let $\mathbb{L}$ be a field and α be algebraic over $\mathbb{L}$. Then, an* ***algebraic extension*** *of $\mathbb{L}$ generated by α, denoted $\mathbb{L}(\alpha)$, is the set of elements of the form*

$$f_0 + f_1\alpha + f_1\alpha^2 + \cdots + f_n\alpha^n \tag{3}$$

where $f_i \in \mathbb{L}$.

Trivially, we can observe that if $\alpha \in \mathbb{L}$, then the expansion would not be interesting as a field is closed under its operations. This algebraic extension "puts" α into $\mathbb{L}$ and allows polynomials with roots α to be factored into $(x - \alpha)$ terms while preserving the characteristic of a field. We can extend $\mathbb{L}$ by all of $\mathbb{L}$'s algebraic elements. Hence, all polynomials $f(x) \in \mathbb{L}(x)$ will factor into polynomials of degree 1, or in another term, **completely split**. This field is called the **algebraic closure** of $\mathbb{L}$, denoted $\overline{\mathbb{F}}$. This mirrors the fact that $\overline{\mathbb{R}} = \mathbb{C}$.

Theorem 2.2 *Let $\mathbb{L}$ be a field. Then $\overline{\mathbb{L}}$ is **algebraically closed**, i.e., every polynomial $f(x) \in \overline{\mathbb{L}}[x]$ has roots in $\overline{\mathbb{L}}$.*

This also means the only algebraic extension of $\overline{\mathbb{L}}$ is $\overline{\mathbb{L}}$ itself i.e., $\overline{\overline{\mathbb{L}}} = \overline{\mathbb{L}}$. An example of an algebraically closed field is $\mathbb{C}$.

Definition 2.13 *Let R be a ring and $G = \{g_1, g_2, \ldots, g_n\}$ be a finite group. $R[G]$ is a **group algebra** i.e., a set of formal sums of the form*

$$a_1 g_1 + a_2 g_2 + \cdots + a_n g_n \tag{4}$$

where and $a_i \in R$.

This construction of a group algebra follows from Corollary 2.1.1 using elements of a group G as the indeterminates. The difference is that the indeterminates are combined based on group G. Since group G is closed under its operation, all elements of $R[G]$ can be reduced into that in Equation 4.

2.2 Matrices and Matrix Multiplication

This section covers some basic concepts regarding matrix multiplication. Firstly, we define matrices and matrix multiplication.

Definition 2.14 *Let $\mathbb{K}$ be a ring and m, n be positive integers. A $m \times n$ **matrix** is a function $a : [m] \times [n] \to \mathbb{K}$. For shorthand notation, define $a_{i,j} := a(i, j)$.*

Definition 2.15 *Let $\mathbb{K}$ be a ring and n be a positive integer. The set of all matrices $a : [n] \times [n] \to \mathbb{K}$ is defined as the **matrix ring** $\mathcal{M}_n(\mathbb{K})$.*

We can verify that $\mathcal{M}_n(\mathbb{K})$ is indeed a ring. Informally, we can refer to a matrix as a rectangular array of elements of R arranged into rows and columns. An example of a 2×3 matrix with entries in $\mathbb{Z}$ is

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad \text{or} \quad a : [2] \times [3] \to \mathbb{Z}, \quad a(i, j) = (i - 1) * 3 + j. \tag{5}$$

Definition 2.16 *Let $\mathbb{K}$ be a ring and $a : [m] \times [n] \to \mathbb{K}$ and $b : [n] \times [p] \to \mathbb{K}$ be matrices. Then,* ***matrix multiplication*** *of a and b is the function $a \cdot b : [m] \times [p] \to \mathbb{K}$ where*

$$(a \cdot b)_{i,j} = \sum_{k=1}^{n} a(i,k)b(k,j) \tag{6}$$

This recovers the same matrix multiplication formula given in linear algebra.

$$\begin{bmatrix} a_{1,1} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \dots & a_{m,n} \end{bmatrix} \cdot \begin{bmatrix} b_{1,1} & \dots & b_{1,p} \\ \vdots & \ddots & \vdots \\ b_{n,1} & \dots & b_{n,p} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{n} a_{1,k}b_{k,1} & \dots & \sum_{k=1}^{n} a_{1,k}b_{k,p} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} a_{m,k}b_{k,1} & \dots & \sum_{k=1}^{n} a_{m,k}b_{k,p} \end{bmatrix} \tag{7}$$

2.2.1 Exponent ω of Matrix Multiplication

We now discuss the exponent ω that researchers use to gauge the efficiency of a matrix multiplication algorithm.

Definition 2.17 *Let $\mathbb{K}$ be a field. Define*

$$M_{\mathbb{K}}(n) := \textit{the number of multiplication operations over } \mathbb{K} \textit{ to multiply two } n \times n \textit{ matrices.} \tag{8}$$

and

$$\omega(\mathbb{K}) := \inf\{\tau \in \mathbb{R} : M_{\mathbb{K}}(n) = \mathcal{O}(n^{\tau})\} \tag{9}$$

Informally, $\omega(\mathbb{K})$ is the minimum exponent required to multiply two $n \times n$ matrices in field $\mathbb{K}$ [2].

Proposition 1 *(Schönhage) If $\mathbb{K}$ is a field extension of $\mathbb{L}$, then $\omega(\mathbb{L}) = \omega(\mathbb{K})$.*

This means $\omega(\mathbb{K})$ is immutable by field extension or restriction. Hence, for this project, we define $\omega := \omega(\mathbb{C})$. Trivially, Equation 7 yields $\omega \leq 3$. It is commonly conjectured that $\omega = 2$.

2.3 Tensors

This section covered some basic information about tensors and how tensors relate to ω. Generally, tensors correspond with a set of computations including matrix multiplication is bilinear. Hence, the discussion of tensors naturally showed up in this project.

2.3.1 Basics

Definition 2.18 *Let $\mathbb{K}$ be a field, $X = \{x_1, x_2, \ldots, x_m\}$, $Y = \{y_1, y_2, \ldots, y_n\}$, and $Z = \{z_1, z_2, \ldots, z_p\}$ be formal variables. A $\mathbb{K}$-**tensor** t over X, Y, X is the formal sum*

$$t = \sum_{i,j,k} t_{i,j,k} x_i y_j z_k \tag{10}$$

where $x_i \in X$, $y_i \in Y$, and $z_i \in Z$ and $t_{i,j,k} \in \mathbb{K}$. We denote $t(x_i, y_j, z_k) = t_{i,j,k}$.

For ease of notation, we will assume a field $\mathbb{K}$ for every tensor if not stated otherwise. Then, we define tensor addition and tensor product. These are not rigorous, but for the sake of this project, this definition is sufficient.

Definition 2.19 *Let $X = \{x_1, x_2, \ldots, x_m\}$, $Y = \{y_1, y_2, \ldots, y_n\}$, and $Z = \{z_1, z_2, \ldots, z_p\}$ be formal variables. Let t and t' be tensors over X, Y, Z, then*

$$t + t' = \sum_{i,j,k} (t_{i,j,k} + t'_{i,j,k})\ x_i y_j z_k \tag{11}$$

is a tensor over X, Y, Z.

Definition 2.20 *Let $X = \{x_1, x_2, \ldots, x_m\}$, $Y = \{y_1, y_2, \ldots, y_n\}$, $Z = \{z_1, z_2, \ldots, z_p\}$, be formal variables. Define X', Y' and Z' similarly. Let t be a tensor over X, Y, Z and t' be a tensor over X', Y', Z', then*

$$t \times t' = \sum_{i,j,k,i',j',k'} (t_{i,j,k} \cdot t'_{i',j',k'})(x_i y_j z_k \cdot x'_i y'_j z'_k). \tag{12}$$

We assume $(x_i y_j z_k \cdot x'_i y'_j z'_k)$ commutes which means we can collect the x, y and z terms. Then,

$$t \times t' = \sum_{i,j,k,i',j',k'} (t_{i,j,k} \cdot t'_{i',j',k'})(x_i x'_i)(y_j y'_j)(z_k z'_k) \tag{13}$$

is a tensor over $X \times X'$, $Y \times Y'$, $Z \times Z'$.

Then, we define what is a matrix multiplication tensor for multiplying $m \times n$ and $n \times p$ matrices.

Definition 2.21 *Let $X = \{x_{i,j}\}_{(i,j) \in [m] \times [n]}$, $Y = \{y_{j,k}\}_{(j,k) \in [n] \times [p]}$, and $Z = \{z_{i,k}\}_{(i,k) \in [m] \times [p]}$ be sets of formal variables. A **matrix multiplication tensor** $\langle m, n, p \rangle$ is given by*

$$\langle m, n, p \rangle := \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} x_{i,j} y_{j,k} z_{i,k}. \tag{14}$$

As the naming suggests, this tensor is equivalent to matrix multiplication. Since we define tensors over formal variables, we can rename the formal variables through a bijection and still get the same tensor. We also have the following proposition.

Proposition 2
$$\prod_{i=1}^{s} \langle m_i, n_i, p_i \rangle = \langle \prod_{i=1}^{s} m_i, \prod_{i=1}^{s} n_i, \prod_{i=1}^{s} p_i \rangle \tag{15}$$

This can be proven by multiplying out all the tensors and collecting the x, y, and z terms in each individual tensor as formal products.

2.3.2 Tensor Rank and Asymptotic Tensor Rank

Definition 2.22 *Let $X = \{x_1, x_2, \ldots, x_m\}$, $Y = \{y_1, y_2, \ldots, y_n\}$, and $Z = \{z_1, z_2, \ldots, y_p\}$, and t be a tensor over X, Y, Z. The* ***rank*** *of tensor t, denoted $R(t)$, is the minimum r such that*

$$t = \sum_{s=1}^{r} \left(\sum_{i=1}^{n} \alpha_{s,i} x_i \right) \left(\sum_{j=1}^{m} \beta_{s,j} y_j \right) \left(\sum_{k=1}^{p} \gamma_{s,k} z_k \right) \tag{16}$$

for some $\alpha_i, \beta_j, \gamma_j \in \mathbb{K}$ and $x_i \in X$, $y_j \in Y$, and $z_i \in Z$.

By definition, if we can write t as $\sum_{s=1}^{r} \left(\sum_{i=1}^{n} \alpha_i x_i \right) \left(\sum_{j=1}^{m} \beta_j y_j \right) \left(\sum_{k=1}^{p} \gamma_k z_k \right)$, $R(t) \leq r$.
In 1969, Strassen [18] related the ω to the rank of a matrix multiplication tensor. Strassen's matrix multiplication algorithm is equivalent to the tensor

$$\begin{aligned} \langle 2,2,2 \rangle = {} & (x_{11} + x_{22})(y_{11} + y_{22})(z_{11} + z_{22}) + (x_{21} + x_{22})y_{11}(z_{21} - z_{22}) \\ & + x_{11}(y_{12} - y_{22})(z_{11} + z_{22}) + x_{22}(y_{21} - y_{11})(z_{11} + z_{21}) + (x_{11} + x_{12})y_{22}(-z_{11} + z_{12}) \\ & + (x_{21} - x_{11})(y_{11} + y_{12})z_{22} + (x_{12} - x_{22})(y_{21} + y_{22})z_{11} \end{aligned} \tag{17}$$

showing $R(\langle 2,2,2 \rangle) \leq 7$, giving $\omega \leq \log_2(7) \approx 2.81$. Winograd [21] later showed that $R(\langle 2,2,2 \rangle) = 7$, giving a tight lowerbound to how fast we can multiply 2×2 matrices.

Proposition 3 *Let t_1 and t_2 be tensors over X_1, Y_1, Z_1 and X_2, Y_2, Z_2, respectively. Then,*

$$R(t_1 + t_2) \geq R(t_1) + R(t_2) \quad \textit{and} \quad R(t_1 t_2) \geq R(t_1) R(t_2) \tag{18}$$

This means that the rank is subadditive and submultiplicative. In general, Proposition 4 shows how the rank of matrix multiplication tensor is related to ω.

Proposition 4

$$\omega := \inf\{\tau \in \mathbb{R} : R(\langle n, n, n\rangle) = \mathcal{O}(n^{\tau})\} \tag{19}$$

The proof for Proposition 4 is available in Bürgisser et al. [7] and Anderson and Barman [2]. The following proposition relates ω to the rank of matrix multiplication tensors of non-square matrices.

Proposition 5 *If $R(\langle m, n, p\rangle) \leq r$, then $(mnp)^{\omega/3} \leq r$ for positive integers m, n, p, and r.*

However, computing the rank of $\langle m, n, p\rangle$ is computationally hard, thus, making the problem of determining ω through the rank difficult. Here we will discuss another type of rank that lets us make use of the definition of ω.

Definition 2.23 *Let t be a tensor over X, Y, Z. The the* ***asymptotic rank*** *of t, denoted $\tilde{R}(t)$, is defined as*

$$\tilde{R}(t) := \lim_{n\to\infty} R(t^n)^{1/n} \tag{20}$$

This limit is well-defined. Similar to rank, the asymptotic rank is also subadditive and submultiplicative. In general, $\tilde{R}(t) \leq R(t)$ for any tensor t.

Proposition 6

$$\tilde{R}(\langle m, n, p\rangle) = (mnp)^{\omega/3} \tag{21}$$

A quick intuition is that while we do not know the base of the best matrix multiplication algorithm with runtime $\mathcal{O}(n^{\omega})$, the actual matrix we are considering is arbitrarily large by definition of the asymptomatic rank, so we can assume we can multiply this arbitrarily large matrix in $\mathcal{O}(n^{\omega})$ time. Then, we define what it is meant for 2 tensors to be the ”same”.

Definition 2.24 *Let t be a tensor over X, Y, Z and t' be a tensor over X', Y', Z'. Then, t and t' are* ***isomorphic*** *if there exists ring isomorphisms $\alpha : \mathbb{K}[X] \to \mathbb{K}[X']$, $\beta : \mathbb{K}[Y] \to \mathbb{K}[Y']$, and $\gamma : \mathbb{K}[Z] \to \mathbb{K}[Z']$ such that $t(x, y, z) = t'(\alpha(x), \beta(y), \gamma(z))$.*

Proposition 7 *Let t be a tensor over X, Y, Z and t' be a tensor over X', Y', Z'. If t and t' are* ***isomorphic*** *then*

$$R(t) = R(t') \quad \textit{and} \quad \tilde{R}(t) = \tilde{R}(t'). \tag{22}$$

However, deriving the asymptomatic rank is even harder than deriving the actual rank of the tensor, which is computationally hard. This leads us to the next section.

2.3.3 Tensor Restriction and Combinatorial Restriction

Definition 2.25 *Let t and t' be tensors defined over X, Y, Z and X', Y', Z'. t is a **restriction** of t' (denoted $t \leq t'$) if there exists ring homomorphisms $\alpha : \mathbb{K}[X] \to \mathbb{K}[X']$, $\beta : \mathbb{K}[Y] \to \mathbb{K}[Y']$, and $\gamma : \mathbb{K}[Z] \to \mathbb{K}[Z']$ such that $t(x, y, z) = t'(\alpha(x), \beta(y), \gamma(z))$.*

Proposition 8 *Let t and t' be tensors defined over X, Y, Z and X', Y', Z'. If $t \leq t'$, then*

$$R(t) \leq R(t') \quad \textit{and} \quad \tilde{R}(t) \leq \tilde{R}(t'). \tag{23}$$

Corollary 2.2.1 *If $\langle m, n, p \rangle \leq t$, $(mnp)^{\omega/3} \leq R(t)$.*

Corollary 2.2.1 is a natural consequence of Proposition 5 and Proposition 8. This gives us some way to give bounds to ω through some tensors whose rank is known. However, the problem of computing the rank of an arbitrary tensor remains hard.

Definition 2.26 *Let t be a tensor defined over X, Y, Z. Let $X' \subseteq X$, $Y' \subseteq Y$, and $Z' \subseteq Z$. Let t' be a tensor defined over X', Y', Z' s.t.,*

$$t'(x, y, z) = \begin{cases} t(x, y, z) & \textit{if } (x, y, z) \in X' \times Y' \times Z' \\ 0 & \textit{otherwise}. \end{cases} \tag{24}$$

*Then, t' is a **combinatorial restriction** (or **zeroing out**) of t.*

Trivially, if t' is a combinatorial restriction (or zeroing out) of t, $t' \leq t$ via natural projections α, β, and γ.

2.3.4 General Method of Bounding ω

By Proposition 3, we have that

$$\sum_{i=1}^{s} R(\langle m_i, n_i, p_i \rangle) \leq R\left(\sum_{i=1}^{s} \langle m_i, n_i, p_i \rangle \right). \tag{25}$$

Then, by applying Proposition 5, we have that

$$\sum_{i=1}^{s} (m_i n_i r_i)^{\omega/3} \leq R\left(\sum_{i=1}^{s} \langle m_i, n_i, p_i \rangle \right). \tag{26}$$

Moreover, if we know an upper bound r to the RHS, we obtain the following theorem.

Theorem 2.3

$$R\left(\sum_{i=1}^{s} \langle m_i, n_i, p_i \rangle \right) \leq r \implies \sum_{i=1}^{s} (m_i n_i p_i)^{\omega/3} \leq r. \tag{27}$$

Similarly, we can do the same thing for the asymptotic rank.

Theorem 2.4 *[19, Section 3]*

$$\tilde{R}\bigg(\sum_{i=1}^{s}\langle m_i, n_i, p_i\rangle\bigg) \leq r \implies \sum_{i=1}^{s}(m_i n_i p_i)^{\omega/3} \leq r. \tag{28}$$

To obtain r, we rely on restricting some tensors with known rank into a sum of matrix multiplication tensors. Since m_i, n_i, p_i, and r are known, we can solve the inequality for an upperbound on ω.
However, for most tensors, deriving for rank is NP-Hard [16]. Thus, in practice, many works [10] rely on the notion of **border rank** and **degeneration** instead of plain rank and restriction which enables more families of tensors can be used. These, however, are not part of this project. More details are described in Bürgisser et al. [7] (Chapters 14 and 15).

2.4 Wedderburn-Artin Theorem

Here are the last ingredients to Cohn & Umans' matrix multiplication. This report will not discuss what it is meant to be **semisimple**, but the structure of this framework guarantees this property.

Theorem 2.5 *(Wedderburn-Artin) Let $\mathcal{A} = \mathbb{K}[G]$ be a **semisimple** group algebra. Then,*

$$\mathcal{A} \cong \mathcal{M}_{d_1}(D_1) \times \mathcal{M}_{d_2}(D_2) \times \mathcal{M}_{d_3}(D_3) \times \cdots \times \mathcal{M}_{d_k}(D_k) \tag{29}$$

*where d_i's are the **character degrees** of G, and d_i and D_i are uniquely determined by G up to permutations, and D_i's are division rings over $\mathbb{K}$. Addition and multiplication in the RHS are pointwise addition and multiplication in the respective matrix rings.*

Theorem 2.6 *If $\mathbb{K}$ is a field and G is a group, then the group algebra $\mathbb{K}[G]$ is semisimple.*

Since we are working or $\mathbb{C}$ (a field) and G in Cohn &Umans' method will always be a group, we are guaranteed Theorem 2.5, and for all i, $D_i = \mathbb{C}$ since $\mathbb{C}$ is algebraically closed. Thus, we have the following corollary.

Corollary 2.6.1 *Let G be a group with character degrees $\{d_i\}$. Then,*

$$\mathbb{C}[G] \cong \mathcal{M}_{d_1}(\mathbb{C}) \times \mathcal{M}_{d_2}(\mathbb{C}) \times \mathcal{M}_{d_3}(\mathbb{C}) \times \cdots \times \mathcal{M}_{d_k}(\mathbb{C}). \tag{30}$$

The following proposition gives us a restriction on what values d_i could be.

Proposition 9 *Let G be a group and $\{d_i\}$ be the character degrees of G. Then, $|G| = \sum d_i^2$.*

3 Cohn and Umans' Matrix Multiplication Framework

This section recounts two approaches to upper bounding ω using a group proposed by Cohn & Umans [8, 9].

Definition 3.1 *Let G be a group and $\mathbb{K}$ be a field. Let $X_G = \{x_g : g \in G\}$, $Y_G = \{y_g : g \in G\}$, $Z_G = \{z_g : g \in G\}$. Then,*

$$T_{\mathbb{K}[G]} := \sum_{g,h \in G} 1_{\mathbb{K}} \cdot x_g y_h z_{gh} \tag{31}$$

*be the **structural tensor** of $\mathbb{K}[G]$.*

The structure tensor of $\mathbb{K}[G]$ described the multiplication of elements in the same group algebra. In both of Cohn & Umans' constructions, the main idea is to reduce the structure tensor of $\mathbb{C}[G]$ into matrix multiplication tensor(s). Since we are only considering $\mathbb{K} = \mathbb{C}$, for ease of notation, we write T_G for the structure tensor of $\mathbb{C}[G]$.

3.1 Triple Product Property (TPP) Construction

The TPP construction aims to restrict this tensor into a matrix multiplication tensor.

Definition 3.2 *Let G be a group and $S \subseteq G$. We define the right quotient set $Q(S)$ as*

$$Q(S) := \{s_1 s_2^{-1} : s_1, s_2 \in S\} \tag{32}$$

Definition 3.3 *[8, Definition 2.1] A group G realizes $\langle m, n, p\rangle$ if there exists subsets S, T, U of G such that $|S| = m$, $|T| = n$, and $|U| = p$, and for all $s \in Q(S)$, $t \in Q(T)$, and $u \in Q(U)$,*

$$stu = 1 \implies s = t = u = 1. \tag{33}$$

*We call this condition of S, T, U the **triple product property**. We say G realizes $\langle m, n, p\rangle$ through S, T, U.*

Indeed, the triple product properties enable matrix multiplication through the following embedding into $\mathbb{C}[G]$:

Let $a : [m] \times [n] \to \mathbb{C}$ and $b : [n] \times [p] \to \mathbb{C}$ be matrices and admit an ordering of $S - \{s_1, \ldots, s_m\}$, $T = \{t_1, \ldots, t_n\}$, $U = \{u_1, \ldots, u_p\}$, then to compute $a \cdot b$, we embed a and b as

$$\bar{A} = \sum_{(i,j)\in[m]\times[n]} a_{i,j} s_i^{-1} t_j \quad \text{and} \quad \bar{B} = \sum_{(j,k)\in[n]\times[p]} b_{j,k} t_j^{-1} u_k. \tag{34}$$

Then, we can read off the coefficients of matrix $c : [m] \times [p] \to \mathbb{C}$ embed into $\mathbb{C}[G]$ as

$$\bar{C} = \bar{A}\bar{B} = \sum_{(i,k)\in[m]\times[p]} c_{i,k} s_i^{-1} u_k. \tag{35}$$

This means $\langle m, n, p\rangle \leq T_G$ i.e., we can combinatorially restrict T_G to obtain matrix multiplication tensor $\langle m, n, p\rangle$. By Corollary 2.6.1,

$$\mathbb{C}[G] \cong \mathcal{M}_{d_1}(\mathbb{C}) \times \mathcal{M}_{d_2}(\mathbb{C}) \times \mathcal{M}_{d_3}(\mathbb{C}) \times \cdots \times \mathcal{M}_{d_k}(\mathbb{C}). \tag{36}$$

Hence, their structural tensors are isomorphic. Hence, by Proposition 7, the structural tensor of the RHS has the same asymptomatic rank as T_G. However, the structural tensor of the RHS is the structural tensor of multiplying square matrices of sizes d_i's. This means

$$\tilde{R}(\langle m, n, p) = (mnp)^{\omega/3} \leq \tilde{R}(T_G) = \sum_{i=1}^{k} \tilde{R}(\langle d_i, d_i, d_i\rangle) = \sum_{i=1}^{k} d_i^{\omega}. \tag{37}$$

Cohn & Umans' TPP construction generates an upperbound for ω through the inequality

$$(mnp)^{\omega/3} \leq \sum_{i=1}^{k} d_i^{\omega} \tag{38}$$

for some group G that realizes $\langle m, n, p\rangle$. Note that by producing the associated mapping for Equation 36 [12], we obtain a concrete recursive matrix multiplication algorithm with runtime corresponding to Equation 38. Since $\{d_i\}$ are constants for a specific group G, we obtain a better matrix multiplication algorithm if we find larger S, T, U in G. In fact, if there exists a TPP construction s.t., $(|S||T||U|)^{2/3} = |G|$, then $\omega = 2$. We can quickly verify this by referring back to Proposition 9.

However, finding the subsets S, T, U in a group G that satisfied the triple product property is not an easy task. Therefore, to facilitate findings components for their framework, Cohn et al. [9] proposed a mathematical object called a Strong Unique Solvable Puzzle (SUSP). Informally, a SUSP can be viewed as a

$s \times k$ matrix with entries in $\{1, 2, 3\}$ satisfying certain properties. Each SUSP of size s and width q with n_i' i entries ($i \in \{1, 2, 3\}$) along with a parameter m implies a group G and its subsets S, T, U of size

$$|G| = s! \cdot m^{sq}, \quad |S| = s! \cdot (m-1)^{n_1}, \quad |T| = s! \cdot (m-1)^{n_2}, \quad |U| = s! \cdot (m-1)^{n_3}. \tag{39}$$

This construction also implies that the maximum character degree of G is $s!$. Generally, larger SUSPs imply a better upper bound of ω. Searching and verifying these SUSPs has been a research focus of Anderson et al. [3] and Anderson and Le [4].

3.2 Simultaneous Triple Product Property (STPP) Construction

Following a similar idea to the TPP construction, the STPP construction aims to realize multiple matrix multiplication tensors through a single group G. Similar to Equation 38, by Theorem 2.4, we obtain ω through the inequality

$$\sum (m_i n_i p_i)^{\omega/3} \leq \sum_{i=1}^{k} d_i^{\omega} \tag{40}$$

where

$$\sum \langle m_i, n_i, p_i \rangle \leq T_G \cong \sum_{i=1}^{k} \langle d_i, d_i, d_i \rangle. \tag{41}$$

This is another attempt to obtain $\omega = 2$ since this results in using more parts of the group G. Rather than using 3 subsets of G, the STTP construction uses multiple triplets of subsets of G.

Definition 3.4 *[9] We says l triplet of subsets S_i, T_i, U_i where $|S_i| = m_i$, $|T_i| = n_i$, $|U_i| = p_i$ and $i \in [l]$ of a group G satisfy the simultaneous triple product property if*

- *S_i, U_i, T_i satisfy the triple product property for every $i \in [l]$.*
- *For every $(i, j, k) \in [l]^3$,*

$$s_i(s_j)^{-1} t_j (t_k)^{-1} u_k (u_i)^{-1} = 1 \implies i = j = k, \tag{42}$$

where $s_i \in S_i$, $s_j \in S_j$, $t_j \in T_j$, $t_k \in T_k$, $u_k \in U_k$, and $u_i \in U_i$.

Then, we say the group G simultaneously realizes $\langle m_1, n_1, p_1 \rangle, \ldots, \langle m_l, n_l, p_l \rangle$.

This construction gives us a bound on ω by Equation 40. Since this construction realizes several matrix multiplications simultaneously, this method is only practical to obtain theoretical bounds on ω. We observe this construction subsumes the TPP construction, which will potentially give better bounds on ω.

4 Methodology

Our research aims to first give a lower bound greater than 2 to all upper bounds on ω given by the TPP construction and the STPP construction, if possible. This section describes a potential approach to the lower bound Cohn & Umans' method.

4.1 Tricolored Sum-Free Sets

Definition 4.1 *Let S, T, U be sets. A* ***perfect matching*** *M of S, T, U is a subset of $S \times T \times U$ s.t., the projection maps $(s, t, u) \mapsto s$, $(s, t, u) \mapsto t$, $(s, t, u) \mapsto u$ are bijective.*

Definition 4.2 *Let $(G, +)$ be a group. A* ***tricolored sum-free set*** *in G is a triple of subsets $S, T, U \subseteq G$ s.t.,*

$$M = \{(s, t, u) \in S \times T \times U : s + t + u = 0\} \tag{43}$$

is a perfect matching in S, T, U.

Note that 0 here is the identity of group G under $+$. If the operation is defined to be multiplicative $*$, an alternative name for the tricolored sum-free set is the tricolored product-free set. Intuitively, we can see some connection between the STTP construction and the tricolored sum-free set.

4.2 Proposed Approach

Blasiak et al. [5, Section 3] has shown a direct connection between the Cohn & Umans' STTP construction and the size of a tricolored sum-free set. Therefore, firstly, to achieve the proposed goal, we need a tight upperbound on the size of tricolored sum-free sets in a group G under some parameters. However, Blasiak et al. [5], Sawin [17], and Blasiak et al. [6] mainly attacked the conjecture achieving $\omega = 2$ with a fixed group G. This does not rule out the possibility of a family of matrix multiplication algorithms with runtimes approaching 2.

Here, I proposed to follow Alman and Williams [1]'s approach where they consider all combinatorial degenerations (not a part of this project) of the structural tensors of $\mathbb{Z}_q$ to derive the best possible upper bound of ω for fixed q. Then, to derive a lower bound, they take $q \to \infty$ to see what is the minimal value.

Similarly, the rough outline for lower bounding Cohn & Umans' method would be considering all combinatorial restrictions of a family of groups into matrix multiplication tensors. Then, through this analysis, we got an upperbound of ω based on a set of parameters. Then, to get the minimal bound on ω derived from Cohn & Umans' method, we take a limit to determine the lower bound. However, a crucial ingredient in this analysis is a tight upper bound on tricolored sum-free sets in a group G, which remains a challenge of this project.

5 Future Plan and Conclusion

For the remaining 10 weeks, I plan to determine which characteristics are affecting the size of the tricolored sum-free set in a group. I plan to do this firstly for groups generated by SUSPs, then a general group. For the latter case, if it is proven to be too ambitious, I would try to reduce the tricolored sum-free set into a different construction that only corresponds to TPP construction and try to redo the analysis under this constraint relaxation. The detailed plan for the next 10 weeks is as follows:

- Week 1-2: Read more research on the tricolored sum-free set.
- Week 3-4: Replicate Alman and Williams [1]'s argument under the background found in Week 1-2 on groups generated by SUSPs (which are TPP constructions). Readjust the project if needed.
- Week 4-5: Proceed with all groups under some characteristics. Readjust the project if needed.
- Week 7-10: Finishing up. Prepare for the poster session, thesis defense, and final report.

Note some adjustments are likely inevitable due to the nature of the project. In my opinion, this plan is reasonable enough for the next term. In fact, if I can solve the latter case, the solution applies to the general matrix multiplication, not only Cohn & Umans' method, so for now, I only hope to be able to solve this problem for some special groups related to Cohn & Umans' matrix multiplication. Since I spent most of the term reading the relevant textbooks and research papers, not much progress has been made towards actually giving a lower bound for the Cohn & Umans' method.

References

[1] Josh Alman and Virginia Vassilevska Williams. "Further Limitations of the Known Approaches for Matrix Multiplication". In: *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Ed. by Anna R. Karlin. Vol. 94. Dagstuhl, Germany, 2018, 25:1–25:15. ISBN: 978-3-95977-060-6. DOI: 10.4230/LIPIcs.ITCS.2018.25.

[2] Matthew Anderson and Siddharth Barman. *The Coppersmith-Winograd Matrix Multiplication Algorithm*. 2009. URL: https://www-auth.cs.wisc.edu/lists/theory-reading/2009-December/pdfmN6UVeUiJ3.pdf.

[3] Matthew Anderson, Zongliang Ji, and Anthony Xu. “Matrix Multiplication: Verifying Strong Uniquely Solvable Puzzles”. In: June 2020, pp. 464–480. ISBN: 978-3-030-51824-0. DOI: 10.1007/978-3-030-51825-7_32.

[4] Matthew Anderson and Vu Le. *Efficiently-Verifiable Strong Uniquely Solvable Puzzles and Matrix Multiplication*. 2023. eprint: arXiv:2307.06463.

[5] Jonah Blasiak et al. “On cap sets and the group-theoretic approach to matrix multiplication”. In: (Aug. 2023). DOI: 10.19086/da.1245.

[6] Jonah Blasiak et al. “Which groups are amenable to proving exponent two for matrix multiplication?” In: *ArXiv* abs/1712.02302 (2017). URL: https://api.semanticscholar.org/CorpusID:24486404.

[7] Peter Bürgisser, Michael Clausen, and Mohammad A. Shokrollahi. *Algebraic Complexity Theory*. 1st. Springer Publishing Company, Incorporated, 2010. ISBN: 3642082289.

[8] Henry Cohn and Christopher Umans. “A group-theoretic approach to fast matrix multiplication”. In: *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.* (2003), pp. 438–449.

[9] Henry Cohn et al. “Group-theoretic algorithms for matrix multiplication”. In: *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)* (2005), pp. 379–388.

[10] Don Coppersmith and Shmuel Winograd. “Matrix multiplication via arithmetic progressions”. In: *Journal of Symbolic Computation* 9.3 (1990). Computational algebraic complexity editorial, pp. 251–280. ISSN: 0747-7171. DOI: https://doi.org/10.1016/S0747-7171(08)80013-2. URL: https://www.sciencedirect.com/science/article/pii/S0747717108800132.

[11] Thomas H. Cormen et al. *Introduction to Algorithms*. 2nd ed. The MIT Press, 2001.

[12] Khai Dong. *Cohn-Umans Matrix Multiplication and Wedderburn Decomposition*. 2023.

[13] Ran Duan, Hongxun Wu, and Renfei Zhou. “Faster Matrix Multiplication via Asymmetric Hashing”. In: *arXiv e-prints*, arXiv:2210.10173 (Oct. 2022), arXiv:2210.10173. DOI: 10.48550/arXiv.2210.10173. arXiv: 2210.10173 [cs.DS].

[14] Zachary Dubinsky. *Fast Matrix Multiplication and The Wedderburn-Artin Theorem*. 2023.

[15] David S. Dummit and Richard M. Foote. *Abstract Algebra*. en. 3rd ed. Nashville, TN: John Wiley Sons, 2003. ISBN: 9780471433347.

[16] Johan Håstad. “Tensor rank is NP-complete”. In: *Journal of Algorithms* 11.4 (1990), pp. 644–654. ISSN: 0196-6774. DOI: https://doi.org/10.1016/0196-6774(90)90014-6. URL: https://www.sciencedirect.com/science/article/pii/0196677490900146.

[17] Will Sawin. “Bounds for Matchings in Nonabelian Groups”. In: *Electron. J. Comb.* 25 (2017), p. 4. URL: https://api.semanticscholar.org/CorpusID:54446488.

[18] V. Strassen. “Gaussian Elimination is not Optimal.” In: *Numerische Mathematik* 13 (1969), pp. 354–356. URL: http://eudml.org/doc/131927.

[19] Virginia Vassilevska Williams. “Limits on All Known (and Some Unknown) Approaches to Matrix Multiplication”. In: *Proceedings of the 2019 on International Symposium on Symbolic and Algebraic Computation*. ISSAC ’19. Beijing, China: Association for Computing Machinery, 2019, p. 10. ISBN: 9781450360845. DOI: 10.1145/3326229.3326282. URL: https://doi.org/10.1145/3326229.3326282.

[20] Virginia Vassilevska Williams et al. “New Bounds for Matrix Multiplication: from Alpha to Omega”. In: *arXiv e-prints*, arXiv:2307.07970 (July 2023), arXiv:2307.07970. DOI: 10.48550/arXiv.2307.07970. arXiv: 2307.07970 [cs.DS].

[21] S. Winograd. “On multiplication of 2 × 2 matrices”. In: *Linear Algebra and its Applications* 4.4 (1971), pp. 381–388. ISSN: 0024-3795. DOI: https://doi.org/10.1016/0024-3795(71)90009-7. URL: https://www.sciencedirect.com/science/article/pii/0024379571900097.