Complexity of Matrix Multiplication and Bilinear Problems

[Handout for the first two lectures]

François Le Gall
Graduate School of Informatics
Kyoto University
legall@i.kyoto-u.ac.jp

1 Introduction

Algebraic complexity theory is the study of computation using algebraic models. One of the main achievements of this field has been the introduction of methods to prove lower bounds on the computational complexity, in algebraic models of computation, of concrete problems. Another major achievement has been the development of powerful techniques to construct fast algorithms for computational problems with an algebraic structure.

The two first lectures will give an overview of some of the main algorithmic applications of algebraic complexity theory, focusing on the construction of bilinear algorithms for computational problems from linear algebra. Our presentation will be systematically illustrated by showing how these ideas from algebraic complexity theory have been used to design asymptotically fast (although not necessarily practical) algorithms for matrix multiplication, as summarized in Table 1. We will show in particular how the techniques described can be applied to construct algorithms that multiply two $n \times n$ matrices over a field using $O(n^{2.38})$ arithmetic operations, which is the best known upper bound on the asymptotic complexity of square matrix multiplication and was first obtained by Coppersmith and Winograd [3].

Table 1: History of the main improvements on the exponent of square matrix multiplication.

Upper bound	Year	Reference	Notes
$\omega \leq 3$			Trivial algorithm
$\omega < 2.81$	1969	Strassen [11]	
$\omega < 2.79$	1979	Pan [6]	
$\omega < 2.78$	1979	Bini et al. [1]	
$\omega < 2.55$	1981	Schönhage [9]	
$\omega < 2.53$	1981	Pan [7]	Not discussed in the lectures
$\omega < 2.52$	1982	Romani [8]	Not discussed in the lectures
$\omega < 2.50$	1982	Coppersmith and Winograd [2]	Not discussed in the lectures
$\omega < 2.48$	1986	Strassen [12]	Not discussed in the lectures
$\omega < 2.376$	1987	Coppersmith and Winograd [3]	
$\omega < 2.374$	2010	Stothers [10] (see also [4])	
$\omega < 2.3729$	2012	Vassilevska Williams [13]	
$\omega < 2.3728639$	2014	Le Gall [5]	

2 Basics of Bilinear Complexity Theory

2.1 Algebraic complexity and the exponent of matrix multiplication

The computation model considered in algebraic complexity theory corresponds to algebraic circuits where each gate represents an elementary algebraic operation over a given field $\mathbb{F}$: addition, subtraction, multiplication, division of two terms, and multiplication of one term by a constant in the field. The algebraic complexity of a problem is the size (i.e., the number of gates) of the smallest algebraic circuit needed to solve the problem.

For instance, for the task of computing the matrix product of two $n \times n$ matrices $A = (a_{ij})_{1 \leq i,j \leq n}$ and $B = (b_{ij})_{1 \leq i,j \leq n}$ with entries in $\mathbb{F}$, we assume that the $2n^2$ entries a_{ij} and b_{ij} are given as input, and want to compute the value

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \tag{2.1}$$

corresponding to the entry in the i-th row and the j-th column of the product AB, for all $(i,j) \in \{1,\ldots,n\} \times \{1,\ldots,n\}$. We will call $\mathcal{C}(n)$ the algebraic complexity of this problem. The exponent of matrix multiplication is denoted ω and defined as

$$\omega = \inf \Big\{ \alpha \,|\, \mathcal{C}(n) \leq n^{\alpha} \text{ for all large enough } n \Big\}.$$

Obviously, $2 \leq \omega \leq 3$.

2.2 Strassen algorithm

Consider the matrix product of two 2×2 matrices. It is easy to see from the definition (Eq. (2.1)) that this product can be computed using 8 multiplications and 4 additions, which implies $\mathcal{C}(2) \leq 12$. In 1969, Strassen [11] showed how to compute this product using only seven multiplications:

1. Compute the following seven terms:

$$\begin{aligned}
m_1 &= a_{11} * (b_{12} - b_{22}),\\
m_2 &= (a_{11} + a_{12}) * b_{22},\\
m_3 &= (a_{21} + a_{22}) * b_{11},\\
m_4 &= a_{22} * (b_{21} - b_{11}),\\
m_5 &= (a_{11} + a_{22}) * (b_{11} + b_{22}),\\
m_6 &= (a_{12} - a_{22}) * (b_{21} + b_{22}),\\
m_7 &= (a_{11} - a_{21}) * (b_{11} + b_{12}).
\end{aligned}$$

2. Output

$$\begin{aligned}
-m_2 + m_4 + m_5 + m_6 &= c_{11},\\
m_1 + m_2 &= c_{12},\\
m_3 + m_4 &= c_{21},\\
m_1 - m_3 + m_5 - m_7 &= c_{22}.
\end{aligned}$$

While the number of multiplication is decreased to seven, the number of additions increases, which does not lead to any improvement for $\mathcal{C}(2)$. The key point is that Strassen's approach can be used *recursively* to compute the product of two $2^k \times 2^k$ matrices, for any $k \geq 1$. By analyzing the complexity of this recursion, we obtain $\mathcal{C}(2^k) = O(7^k)$, which implies that

$$\omega \leq \log_2(7) = 2.80735...,$$

which was the upper bound obtained in [11].

2.3 Bilinear algorithms

A bilinear algorithm for matrix multiplication is an algebraic algorithm that proceeds in two steps. First, t products of the form

$$\begin{aligned} m_1 &= (\text{linear combination of the } a_{ij}\text{'s}) * (\text{linear combination of the } b_{ij}\text{'s}) \\ &\vdots \\ m_t &= (\text{linear combination of the } a_{ij}\text{'s}) * (\text{linear combination of the } b_{ij}\text{'s}) \end{aligned}$$

are computed. Then, each entry c_{ij} is computed by taking a linear combination of $m_1, \ldots, m_t$. The integer t is called the bilinear complexity of this algorithm. Strassen's algorithm is an example of bilinear algorithm that computes the product of two 2×2 matrices with bilinear complexity $t = 7$.

Strassen's recursive approach can actually be generalized to any bilinear algorithm for matrix multiplication, as stated in the following proposition.

Proposition 1. *Let m be a positive integer. Suppose that there exists a bilinear algorithm that computes the product of two $m \times m$ matrices with bilinear complexity t. Then*

$$\omega \leq \log_m(t).$$

3 First Techniques

We introduce the concepts of tensor and rank. We then use these concepts to obtain Theorem 1 below, which generalizes Proposition 1.

3.1 Tensors

Consider three finite-dimensional vector spaces U, V and W over the field $\mathbb{F}$. Take a basis $\{x_1, \ldots, x_{\dim(U)}\}$ of U, a basis $\{y_1, \ldots, y_{\dim(V)}\}$ of V, and a basis $\{z_1, \ldots, z_{\dim(W)}\}$ of W. A *tensor* over (U, V, W) is an element of $U \otimes V \otimes W$ or, equivalently, a formal sum

$$T = \sum_{u=1}^{\dim U} \sum_{v=1}^{\dim V} \sum_{w=1}^{\dim W} d_{uvw}\, x_u \otimes y_v \otimes z_w$$

with coefficient $d_{uvw} \in \mathbb{F}$ for each $(u, v, w) \in \{1, \ldots, \dim(U)\} \times \{1, \ldots, \dim(V)\} \times \{1, \ldots, \dim(W)\}$.

The tensor corresponding to the multiplication of an $m \times n$ matrix by an $n \times p$ matrix is defined as follows. First take $\dim(U) = mn$, $\dim(V) = np$ and $\dim(W) = mp$. It is convenient to change the notation for the bases and denote $\{a_{ij}\}_{1 \leq i \leq m, 1 \leq j \leq n}$ the basis of U, $\{b_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq p}$ the basis of V and $\{c_{ij}\}_{1 \leq i \leq m, 1 \leq j \leq p}$ the basis of W. In this case, an arbitrary tensor a over (U, V, W) is a formal sum

$$T = \sum_{i,i'=1}^{m} \sum_{k,k'=1}^{n} \sum_{j,j'=1}^{p} d_{ii'jj'kk'}\, a_{ik} \otimes b_{k'j} \otimes c_{i'j'}.$$

The tensor corresponding to the multiplication of an $m \times n$ matrix by an $n \times p$ matrix is the tensor with

$$d_{ii'jj'kk'} = \begin{cases} 1 & \text{if } i = i' \text{ and } j = j' \text{ and } k = k', \\ 0 & \text{otherwise .} \end{cases}$$

We summarize this definition as follows.

Definition 1. *The tensor corresponding to the multiplication of an $m \times n$ matrix by an $n \times p$ matrix is*

$$\sum_{i=1}^{m} \sum_{j=1}^{p} \sum_{k=1}^{n} a_{ik} \otimes b_{kj} \otimes c_{ij}. \tag{3.1}$$

This tensor is denoted $\langle m, n, p \rangle$.

One can define in a natural way the tensor product of two tensors. In particular, for matrix multiplication tensors, we obtain the following identity: for any positive integers m, m', n, n', p, p',

$$\langle m, n, p \rangle \otimes \langle m', n', p' \rangle \cong \langle mm', nn', pp' \rangle. \tag{3.2}$$

3.2 The rank of a tensor

We now define the rank of tensor.

Definition 2. *Let T be a tensor over (U, V, W). The rank of T, denoted $R(T)$, is the minimal integer t for which T can be written as*

$$T = \sum_{s=1}^{t} \left[\left(\sum_{u=1}^{\dim(U)} \alpha_{su} x_u \right) \otimes \left(\sum_{v=1}^{\dim(V)} \beta_{sv} y_v \right) \otimes \left(\sum_{w=1}^{\dim(W)} \gamma_{sw} z_w \right) \right],$$

for some constants $\alpha_{su}, \beta_{sv}, \gamma_{sw}$ in $\mathbb{F}$.

As an illustration of this definition, consider the rank of the matrix multiplication tensor. Obviously, $R(\langle m, n, p \rangle) \leq mnp$, from the definition (Eq. (3.1)). In particular, $R(2, 2, 2) \leq 8$. Strassen's algorithm corresponds to the equality

$$\begin{aligned}
\langle 2, 2, 2 \rangle = & a_{11} \otimes (b_{12} - b_{22}) \otimes (c_{12} + c_{22}) \\
& + (a_{11} + a_{12}) \otimes b_{22} \otimes (-c_{11} + c_{12}) \\
& + (a_{21} + a_{22}) \otimes b_{11} \otimes (c_{21} - c_{22}) \\
& + a_{22} \otimes (b_{21} - b_{11}) \otimes (c_{11} + c_{21}) \\
& + (a_{11} + a_{22}) \otimes (b_{11} + b_{22}) \otimes (c_{11} + c_{22}) \\
& + (a_{12} - a_{22}) \otimes (b_{21} + b_{22}) \otimes c_{11} \\
& + (a_{11} - a_{21}) \otimes (b_{11} + b_{12}) \otimes (-c_{22}),
\end{aligned}$$

which shows that actually $R(2, 2, 2) \leq 7$. More generally, it is easy to see that the rank of a matrix multiplication tensor corresponds to the bilinear complexity of the best bilinear algorithm that computes this matrix multiplication. This observation directly implies the following reinterpretation of Proposition 1: if $R(m, m, m) \leq t$ then $\omega \leq \log_m(t)$. This argument can be generalized as follows.

Theorem 1. *Let m, n, p and t be four positive integers. If $R(\langle m, n, p \rangle) \leq t$, then*

$$(mnp)^{\omega/3} \leq t.$$

This theorem can be proved using the following two properties of the rank. First, for any tensors T and T',

$$R(T \otimes T') \leq R(T) \times R(T'). \tag{3.3}$$

Secondly, for any positive integers m, n and p,

$$R(\langle m, n, p \rangle) = R(\langle m, p, n \rangle) = R(\langle n, m, p \rangle) = R(\langle n, p, m \rangle) = R(\langle p, m, n \rangle) = R(\langle p, n, m \rangle). \tag{3.4}$$

Note that this second property has interesting consequences. It implies, for instance, that the bilinear complexity of computing the product of an $n \times n$ matrix by an $n \times n^2$ matrix is the same as the bilinear complexity of computing the product of an $n \times n^2$ matrix by an $n^2 \times n$.

4 More Advanced Techniques

We introduce the concept of border rank of a tensor, and use it to obtain Theorem 2 below, which generalizes Theorem 1. We then present Schönhage's asymptotic sum inequality (Theorem 3), which significantly generalizes Theorem 2.

4.1 Approximate bilinear algorithms and border rank

Let λ be an indeterminate and $\mathbb{F}[\lambda]$ denote the ring of polynomials in λ with coefficients in the field $\mathbb{F}$. We now define the concept of border rank of a tensor (compare with Definition 2).

Definition 3. *Let T be a tensor over (U, V, W). The border rank of T, denoted $\underline{R}(T)$, is the minimal integer t for which there exist an integer $c \geq 0$ and a tensor T'' such that T can be written as as*

$$\lambda^c T = \sum_{s=1}^{t} \left[\left(\sum_{u=1}^{\dim(U)} \alpha_{su} x_u \right) \otimes \left(\sum_{v=1}^{\dim(V)} \beta_{sv} y_v \right) \otimes \left(\sum_{w=1}^{\dim(W)} \gamma_{sw} z_w \right) \right] + \lambda^{c+1} T'',$$

for some constants $\alpha_{su}, \beta_{sv}, \gamma_{sw}$ in $\mathbb{F}[\lambda]$.

Obviously, $\underline{R}(T) \leq R(T)$ for any tensor T. Moreover, Eqs. (3.3) and (3.4) hold when replacing the rank by the border rank.

Let us study an example. Bini et al. [1] considered the tensor

$$\begin{aligned} T_{\mathsf{Bini}} = & \sum_{\substack{1 \leq i,j,k \leq 2 \\ (i,k) \neq (2,2)}} a_{ik} \otimes b_{kj} \otimes c_{ij} \\ = & a_{11} \otimes b_{11} \otimes c_{11} + a_{12} \otimes b_{21} \otimes c_{11} + a_{11} \otimes b_{12} \otimes c_{12} + a_{12} \otimes b_{22} \otimes c_{12} \\ & + a_{21} \otimes b_{11} \otimes c_{21} + a_{21} \otimes b_{12} \otimes c_{22}, \end{aligned}$$

which corresponds to a matrix product of two 2×2 matrices where one entry in the first matrix is zero (more precisely, $a_{22} = 0$). It can be shown that $R(T_{\mathsf{Bini}}) = 6$. Bini et al. [1] showed that $\underline{R}(T_{\mathsf{Bini}}) \leq 5$ by exhibiting the identity

$$\lambda T_{\mathsf{Bini}} = T' + \lambda^2 T''$$

where

$$\begin{aligned} T' = & (a_{12} + \lambda a_{11}) \otimes (b_{12} + \lambda b_{22}) \otimes c_{12} \\ & + (a_{21} + \lambda a_{11}) \otimes b_{11} \otimes (c_{11} + \lambda c_{21}) \\ & - a_{12} \otimes b_{12} \otimes (c_{11} + c_{12} + \lambda c_{22}) \\ & - a_{21} \otimes (b_{11} + b_{12} + \lambda b_{21}) \otimes c_{11} \\ & + (a_{12} + a_{21}) \otimes (b_{12} + \lambda b_{21}) \otimes (c_{11} + \lambda c_{22}) \end{aligned}$$

and

$$T'' = a_{11} \otimes b_{22} \otimes c_{12} + a_{11} \otimes b_{11} \otimes c_{21} + (a_{12} + a_{21}) \otimes b_{21} \otimes c_{22}.$$

Remember that the rank of a tensor is related to the complexity of bilinear algorithms computing the tensor. The border rank is related to the complexity of *approximate* bilinear algorithms computing the tensor. Another contribution of [1] was to show that approximate bilinear algorithms can be converted into usual bilinear algorithms without increasing the complexity too much, as stated in the following proposition.

Proposition 2. *There exists a constant a such that $R(T) \leq a \times \underline{R}(T)$ for any tensor T.*

The constant a in Proposition 2 actually depends of the value c in the border rank. We will nevertheless ignore this technical point in these lectures.

It is easy to see, by combining two copies of the tensor T_{Bini}, that $\underline{R}(\langle 3,3,2\rangle) \leq 10$. From the border rank versions of Eqs. (3.3) and (3.4), this gives $\underline{R}(\langle 12,12,12\rangle) \leq 1000$, and thus

$$R(\langle 12,12,12\rangle) \leq a \times 1000$$

from Proposition 2. Unfortunately, this inequality (via Theorem 1) does not give any interesting upper bound on ω unless a is very close to one, which is not the case (indeed, in the example we are now studying the constant a can be taken as 10). The trick to bypass this difficulty is to consider the tensor $\langle 12,12,12\rangle^{\otimes N} \cong \langle 12^N, 12^N, 12^N\rangle$ for a large integer N. We have

$$R(\langle 12^N, 12^N, 12^N\rangle) \leq a \times \underline{R}(\langle 12^N, 12^N, 12^N\rangle) \leq a \times 1000^N.$$

Now this inequality, via Theorem 1, gives $\omega \leq \log_{12}(a^{1/N} \times 1000)$, which implies

$$\omega \leq \log_{12}(1000) < 2.78$$

by taking the limit when N goes to infinity. This upper bound was obtained by Bini et al. [1].

The above analysis indicates that, when deriving an upper bound on ω via Theorem 1, one can use the border rank instead of the rank. Indeed, the rank can be replaced by the border rank in Theorem 1, as we now state.

Theorem 2. *Let m, n, p and t be four positive integers. If $\underline{R}(\langle m,n,p\rangle) \leq t$, then*

$$(mnp)^{\omega/3} \leq t.$$

4.2 Schönhage's asymptotic sum inequality

Schönhage [9] considered the following tensor:

$$T_{\mathsf{Schon}} = \sum_{i,j=1}^{3} a_i \otimes b_j \otimes c_{ij} + \sum_{k=1}^{4} v_k \otimes v_k \otimes w.$$

Observe that the first part is isomorphic to $\langle 3,1,3\rangle$, and the second part is isomorphic to $\langle 1,4,1\rangle$. Since the first part and the second part do not share variables, the sum is actually direct, so we have

$$T_{\mathsf{Schon}} \cong \langle 3,1,3\rangle \oplus \langle 1,4,1\rangle.$$

Since $\underline{R}(\langle 3,1,3\rangle) = 9$ and $\underline{R}(\langle 1,4,1\rangle) = 4$, we obtain immediately $\underline{R}(T_{\mathsf{Schon}}) \leq 13$. Schönhage showed that

$$\underline{R}(T_{\mathsf{Schon}}) \leq 10,$$

by exhibiting the identity

$$\lambda^2 T_{\mathsf{Schon}} = T' + \lambda^3 T''$$

for

$$\begin{aligned}T' =&(a_1 + \lambda u_1) \otimes (b_1 + \lambda v_1) \otimes (w + \lambda^2 c_{11}) \\ &+ (a_1 + \lambda u_2) \otimes (b_2 + \lambda v_2) \otimes (w + \lambda^2 c_{12}) \\ &+ (a_2 + \lambda u_3) \otimes (b_1 + \lambda v_3) \otimes (w + \lambda^2 c_{21}) \\ &+ (a_2 + \lambda u_4) \otimes (b_2 + \lambda v_4) \otimes (w + \lambda^2 c_{22}) \\ &+ (a_3 - \lambda u_1 - \lambda u_3) \otimes b_1 \otimes (w + \lambda^2 c_{31}) \\ &+ (a_3 - \lambda u_2 - \lambda u_4) \otimes b_2 \otimes (w + \lambda^2 c_{32}) \\ &+ a_1 \otimes (b_3 - \lambda v_1 - \lambda v_2) \otimes (w + \lambda^2 c_{13}) \\ &+ a_2 \otimes (b_3 - \lambda v_3 - \lambda v_4) \otimes (w + \lambda^2 c_{23}) \\ &+ a_3 \otimes b_3 \otimes (w + \lambda^2 c_{33}) \\ &- (a_1 + a_2 + a_3) \otimes (b_1 + b_2 + b_3) \otimes w\end{aligned}$$

and some tensor T''.

Schönhage also showed the following strong generalization of Theorem 2 known as the *asymptotic sum inequality*.

Theorem 3. *Let k, t be two positive integers, and $m_1, \ldots, m_k$, $n_1, \ldots, n_k$, $p_1, \ldots, p_k$ be $3k$ positive integers. If*

$$\underline{R}\left(\bigoplus_{i=1}^{k}\langle m_i, n_i, p_i\rangle\right) \leq t$$

then

$$\sum_{i=1}^{k}(m_i n_i p_i)^{\omega/3} \leq t.$$

Applying Theorem 3 to T_{Schon} gives

$$9^{\omega/3} + 4^{\omega/3} \leq 10,$$

which implies $\omega \leq 2.60$. Using a variant of this tensor, Schönhage [9] ultimately obtained the upper bound

$$\omega \leq 2.55,$$

again via Theorem 3.

5 The Laser Method

We show how the techniques developed so far, combined with an approach sometimes called the *laser method*, can be applied to obtain the upper bound $\omega < 2.38$. This upper bound has been first obtained by Coppersmith and Winograd [3].

5.1 The first construction by Coppersmith and Winograd

We start with the first construction from Coppersmith and Winograd's paper [3].

Let q be a positive integer, and consider three vector spaces U, V and W of dimension $q + 1$ over the field $\mathbb{F}$. Take a basis $\{x_0, \ldots, x_q\}$ of U, a basis $\{y_0, \ldots, y_q\}$ of V, and a basis $\{z_0, \ldots, z_q\}$ of W. Consider the tensor

$$T_{\mathsf{easy}} = T^{011}_{\mathsf{easy}} + T^{101}_{\mathsf{easy}} + T^{110}_{\mathsf{easy}},$$

where

$$T^{011}_{\mathsf{easy}} = \sum_{i=1}^{q} x_0 \otimes y_i \otimes z_i \cong \langle 1, 1, q \rangle,$$
$$T^{101}_{\mathsf{easy}} = \sum_{i=1}^{q} x_i \otimes y_0 \otimes z_i \cong \langle q, 1, 1 \rangle,$$
$$T^{110}_{\mathsf{easy}} = \sum_{i=1}^{q} x_i \otimes y_i \otimes z_0 \cong \langle 1, q, 1 \rangle.$$

Remark. The superscripts in T^{011}_{easy}, T^{101}_{easy} and T^{110}_{easy} come from the following decomposition of the three vector spaces U, V and W:

$$U = U_0 \oplus U_1 \text{ where } U_0 = span\{x_0\} \text{ and } U_1 = span\{x_1, \ldots, x_q\}$$
$$V = V_0 \oplus V_1, \text{ where } V_0 = span\{y_0\} \text{ and } V_1 = span\{y_1, \ldots, y_q\}$$
$$W = W_0 \oplus W_1, \text{ where } W_0 = span\{z_0\} \text{ and } W_1 = span\{z_1, \ldots, z_q\}.$$

Observe that

$$\lambda^3 T_{\mathsf{easy}} = T' + \lambda^4 T''$$

where

$$\begin{aligned} T' = &\sum_{i=1}^{q} \lambda(x_0 + \lambda x_i) \otimes (y_0 + \lambda y_i) \otimes (z_0 + \lambda z_i) \\ &- (x_0 + \lambda^2 \sum_{i=1}^{q} x_i) \otimes (y_0 + \lambda^2 \sum_{i=1}^{q} y_i) \otimes (z_0 + \lambda^2 \sum_{i=1}^{q} z_i) \\ &+ (1 - q\lambda) x_0 \otimes y_0 \otimes z_0. \end{aligned}$$

and T'' is some tensor. Thus $\underline{R}(T_{\mathsf{easy}}) \leq q + 2$.

While the tensor T_{easy} is a sum of three parts (T^{011}_{easy}, T^{101}_{easy} and T^{110}_{easy}), Theorem 3 cannot be used since the sum is not direct (for instance, the variable y_i, for any $i \in \{1, \ldots, q\}$, is shared by T^{011}_{easy} and T^{110}_{easy}). Coppersmith and Winograd [3] showed how to overcome this difficulty by considering several copies of T_{easy}, and obtained the following result, which is an illustration of the *laser method* developed by Strassen [12].

Theorem 4. *For N large enough, the tensor $T^{\otimes N}_{\mathsf{easy}}$ can be converted into a direct sum of*

$$2^{\left(H\left(\frac{1}{3},\frac{2}{3}\right)-o(1)\right)N}$$

terms[1]*, each isomorphic to*

$$\left[T^{011}_{\mathsf{easy}}\right]^{\otimes N/3} \otimes \left[T^{101}_{\mathsf{easy}}\right]^{\otimes N/3} \otimes \left[T^{110}_{\mathsf{easy}}\right]^{\otimes N/3} \cong \langle q^{N/3}, q^{N/3}, q^{N/3} \rangle.$$

Theorem 4 implies, via Theorem 3, that

$$2^{\left(H\left(\frac{1}{3},\frac{2}{3}\right)-o(1)\right)N} \times q^{N\omega/3} \leq \underline{R}(T^{\otimes N}_{\mathsf{easy}}) \leq (q+2)^N.$$

[1]Here $H\left(\frac{1}{3}, \frac{2}{3}\right) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$ represents the entropy (with logarithms taken to the basis 2) of the corresponding probability distribution.

We thus obtain

$$2^{H(\frac{1}{3},\frac{2}{3})} \times q^{\omega/3} \leq q+2,$$

which gives

$$\omega < 2.41$$

for $q = 8$.

5.2 The second construction by Coppersmith and Winograd

We now describe the second construction from Coppersmith and Winograd's paper [3].

Let q be a positive integer, and consider three vector spaces U, V and W of dimension $q+2$ over $\mathbb{F}$. Take a basis $\{x_0, \ldots, x_{q+1}\}$ of U, a basis $\{y_0, \ldots, y_{q+1}\}$ of V, and a basis $\{z_0, \ldots, z_{q+1}\}$ of W. Consider the tensor

$$T_{\mathsf{CW}} = T_{\mathsf{CW}}^{011} + T_{\mathsf{CW}}^{101} + T_{\mathsf{CW}}^{110} + T_{\mathsf{CW}}^{002} + T_{\mathsf{CW}}^{020} + T_{\mathsf{CW}}^{200},$$

where

$$\begin{aligned}
T_{\mathsf{CW}}^{011} &= \sum_{i=1}^{q} x_0 \otimes y_i \otimes z_i \cong \langle 1,1,q \rangle, \\
T_{\mathsf{CW}}^{101} &= \sum_{i=1}^{q} x_i \otimes y_0 \otimes z_i \cong \langle 1,q,1 \rangle, \\
T_{\mathsf{CW}}^{110} &= \sum_{i=1}^{q} x_i \otimes y_i \otimes z_0 \cong \langle q,1,1 \rangle, \\
T_{\mathsf{CW}}^{002} &= x_0 \otimes y_0 \otimes z_{q+1} \cong \langle 1,1,1 \rangle, \\
T_{\mathsf{CW}}^{020} &= x_0 \otimes y_{q+1} \otimes z_0 \cong \langle 1,1,1 \rangle, \\
T_{\mathsf{CW}}^{200} &= x_{q+1} \otimes y_0 \otimes z_0 \cong \langle 1,1,1 \rangle.
\end{aligned}$$

Remark. The superscripts in T_{CW}^{011}, T_{CW}^{101}, T_{CW}^{110}, T_{CW}^{200}, T_{CW}^{020} and T_{CW}^{002} come from the following decomposition of the three vector spaces U, V and W:

$$\begin{aligned}
&U = U_0 \oplus U_1 \oplus U_2 \text{ where } U_0 = span\{x_0\}, U_1 = span\{x_1, \ldots, x_q\} \text{ and } U_2 = span\{x_{q+1}\} \\
&V = V_0 \oplus V_1 \oplus V_2, \text{ where } V_0 = span\{y_0\}, V_1 = span\{y_1, \ldots, y_q\} \text{ and } V_2 = span\{y_{q+1}\} \\
&W = W_0 \oplus W_1 \oplus W_2, \text{ where } W_0 = span\{z_0\}, W_1 = span\{z_1, \ldots, z_q\} \text{ and } W_2 = span\{z_{q+1}\}.
\end{aligned}$$

Observe that

$$\lambda^3 T_{\mathsf{CW}} = T' + \lambda^4 T''$$

where

$$\begin{aligned}
T' = &\sum_{i=1}^{q} \lambda(x_0 + \lambda x_i) \otimes (y_0 + \lambda y_i) \otimes (z_0 + \lambda z_i) \\
&- (x_0 + \lambda^2 \sum_{i=1}^{q} x_i) \otimes (y_0 + \lambda^2 \sum_{i=1}^{q} y_i) \otimes (z_0 + \lambda^2 \sum_{i=1}^{q} z_i) \\
&+ (1 - q\lambda)(x_0 + \lambda^3 x_{q+1}) \otimes (y_0 + \lambda^3 y_{q+1}) \otimes (z_0 + \lambda^3 z_{q+1}).
\end{aligned}$$

and T'' is some tensor. Thus $\underline{R}(T_{\mathsf{CW}}) \leq q+2$.

While the tensor T_{CW} is a sum of six parts, Theorem 3 cannot directly be used since the sum is not direct. Again, Coppersmith and Winograd [3] showed that this difficulty can be overcome by considering many copies of T_{CW}, and obtained the following result.

Theorem 5. *For any $0 \leq \alpha \leq 1/3$ and for N large enough, the tensor $T_{\mathsf{CW}}^{\otimes N}$ can be converted into a direct sum of*

$$2^{\left(H\left(\frac{2}{3}-\alpha,2\alpha,\frac{1}{3}-\alpha\right)-o(1)\right)N}$$

terms, each isomorphic to

$$\left[T_{\mathsf{CW}}^{011}\right]^{\otimes \alpha N} \otimes \left[T_{\mathsf{CW}}^{101}\right]^{\otimes \alpha N} \otimes \left[T_{\mathsf{CW}}^{110}\right]^{\otimes \alpha N} \otimes \left[T_{\mathsf{CW}}^{002}\right]^{\otimes \left(\frac{1}{3}-\alpha\right)N} \otimes \left[T_{\mathsf{CW}}^{020}\right]^{\otimes \left(\frac{1}{3}-\alpha\right)N} \otimes \left[T_{\mathsf{CW}}^{200}\right]^{\otimes \left(\frac{1}{3}-\alpha\right)N} \cong \langle q^{\alpha N}, q^{\alpha N}, q^{\alpha N}\rangle.$$

Theorem 5 implies, via Theorem 3, that

$$2^{\left(H\left(\frac{2}{3}-\alpha,2\alpha,\frac{1}{3}-\alpha\right)-o(1)\right)N} \times q^{\alpha N \omega} \leq \underline{R}(T_{\mathsf{CW}}^{\otimes N}) \leq (q+2)^N.$$

We thus obtain

$$2^{H\left(\frac{2}{3}-\alpha,2\alpha,\frac{1}{3}-\alpha\right)} \times q^{\alpha\omega} \leq q+2,$$

which gives

$$\omega < 2.3871900$$

for $q = 6$ and $\alpha = 0.3173$.

5.3 Taking powers of the second construction by Coppersmith and Winograd

Consider the tensor

$$T_{\mathsf{CW}}^{\otimes 2} = T_{\mathsf{CW}} \otimes T_{\mathsf{CW}}.$$

We can write

$$\begin{aligned} T_{\mathsf{CW}}^{\otimes 2} = {} & T^{400} + T^{040} + T^{004} + T^{310} + T^{301} + T^{103} + T^{130} + T^{013} + T^{031} + T^{220} + T^{202} + T^{022} \\ & + T^{211} + T^{121} + T^{112}, \end{aligned}$$

where

$$\begin{aligned} T^{400} &= T_{\mathsf{CW}}^{200} \otimes T_{\mathsf{CW}}^{200}, \\ T^{310} &= T_{\mathsf{CW}}^{200} \otimes T_{\mathsf{CW}}^{110} + T_{\mathsf{CW}}^{110} \otimes T_{\mathsf{CW}}^{200}, \\ T^{220} &= T_{\mathsf{CW}}^{200} \otimes T_{\mathsf{CW}}^{020} + T_{\mathsf{CW}}^{020} \otimes T_{\mathsf{CW}}^{200} + T_{\mathsf{CW}}^{110} \otimes T_{\mathsf{CW}}^{110}, \\ T^{211} &= T_{\mathsf{CW}}^{200} \otimes T_{\mathsf{CW}}^{011} + T_{\mathsf{CW}}^{011} \otimes T_{\mathsf{CW}}^{200} + T_{\mathsf{CW}}^{110} \otimes T_{\mathsf{CW}}^{101} + T_{\mathsf{CW}}^{101} \otimes T_{\mathsf{CW}}^{110}, \end{aligned}$$

and the other 11 terms are obtained by permuting the variables (e.g., $T^{040} = T_{\mathsf{CW}}^{020} \otimes T_{\mathsf{CW}}^{020}$).

Coppersmith and Winograd [3] showed how to generalize the approach of Section 5.2 to analyze $T_{\mathsf{CW}}^{\otimes 2}$, and obtained the upper bound

$$\omega \leq 2.3754770$$

by solving an optimization problem of 3 variables (remember that in Section 5.2 the optimization problem had only one variable α).

Since $T_{\mathsf{CW}}^{\otimes 2}$ gives better upper bounds on ω than T_{CW}, a natural question was to consider higher powers of T_{CW}, i.e., study the tensor $T_{\mathsf{CW}}^{\otimes m}$ for $m \geq 3$. Investigating the third power (i.e., $m = 3$) was indeed explicitly mentioned as an open problem in [3]. More that twenty years later, Stothers showed that, while the third power does not seem to lead to any improvement, the fourth power does give an improvement [10]. The cases $m = 8$, $m = 16$ and $m = 32$ have then been analyzed, giving the upper bounds on ω summarized in Table 2.

Table 2: Upper bounds on ω obtained by analyzing the m-th power of the construction T_{CW}.

m	Upper bound	Number of variables in in the optimization problem	Reference
1	$\omega < 2.3871900$	1	Ref. [3]
2	$\omega < 2.3754770$	3	Ref. [3]
4	$\omega < 2.3729269$	9	Ref. [10, 13]
8	$\omega < 2.3728642$	29	Ref. [5] ($\omega < 2.3729$ given in Ref. [13])
16	$\omega < 2.3728640$	101	Ref. [5]
32	$\omega < 2.3728639$	373	Ref. [5]

References

[1] BINI, D., CAPOVANI, M., ROMANI, F., AND LOTTI, G. $O(n^{2.7799})$ complexity for $n \times n$ approximate matrix multiplication. *Information Processing Letters 8*, 5 (1979), 234–235.

[2] COPPERSMITH, D., AND WINOGRAD, S. On the asymptotic complexity of matrix multiplication. *SIAM Journal on Computing 11*, 3 (1982), 472–492.

[3] COPPERSMITH, D., AND WINOGRAD, S. Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation 9*, 3 (1990), 251–280.

[4] DAVIE, A. M., AND STOTHERS, A. J. Improved bound for complexity of matrix multiplication. *Proceedings of the Royal Society of Edinburgh 143A* (2013), 351–370.

[5] LE GALL, F. Powers of tensors and fast matrix multiplication. In *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation* (2014), 296–303.

[6] PAN, V. Y. Field extension and triangular aggregating, uniting and canceling for the acceleration of matrix multiplications. In *Proceedings of the 20th Annual Symposium on Foundations of Computer Science* (1979), pp. 28–38.

[7] PAN, V. Y. New combinations of methods for the acceleration of matrix multiplication. *Computer and Mathematics with Applications* (1981), 73–125.

[8] ROMANI, F. Some properties of disjoint sums of tensors related to matrix multiplication. *SIAM Journal on Computing 11*, 2 (1982), 263–267.

[9] SCHÖNHAGE, A. Partial and total matrix multiplication. *SIAM Journal on Computing 10*, 3 (1981), 434–455.

[10] STOTHERS, A. *On the Complexity of Matrix Multiplication*. PhD thesis, University of Edinburgh, 2010.

[11] STRASSEN, V. Gaussian elimination is not optimal. *Numerische Mathematik 13* (1969), 354–356.

[12] STRASSEN, V. The asymptotic spectrum of tensors and the exponent of matrix multiplication. In *Proceedings of the 27th Annual Symposium on Foundations of Computer Science* (1986), pp. 49–54.

[13] VASSILEVSKA WILLIAMS, V. Multiplying matrices faster than Coppersmith-Winograd. In *Proceedings of the 44th Symposium on Theory of Computing* (2012), pp. 887–898.