Electronic Colloquium on Computational Complexity, Report No. 30 (2020)

Barriers for Rectangular Matrix Multiplication

Matthias Christandl *University of Copenhagen* christandl@math.ku.dk
François Le Gall *Nagoya University* legall@math.nagoya-u.ac.jp
Vladimir Lysikov *University of Copenhagen* vl@math.ku.dk
Jeroen Zuiddam *Institute for Advanced Study* jzuiddam@ias.edu

Abstract. We study the algorithmic problem of multiplying large matrices that are rectangular. We prove that the method that has been used to construct the fastest algorithms for rectangular matrix multiplication cannot give optimal algorithms. In fact, we prove a precise numerical barrier for this method. Our barrier improves the previously known barriers, both in the numerical sense, as well as in its generality. We prove our result using the asymptotic spectrum of tensors. More precisely, we crucially make use of two families of real tensor parameters with special algebraic properties: the quantum functionals and the support functionals. In particular, we prove that any lower bound on the dual exponent of matrix multiplication α via the big Coppersmith–Winograd tensors cannot exceed 0.625.

1. Introduction

Given two large matrices, how many arithmetic operations, plus and times, are required to compute their matrix product?

The high school algorithm for multiplying two square matrices of shape $n \times n$ costs roughly $2n^3$ arithmetic operations. On the other hand, we know that at least n^2 operations are required. Denoting by ω the optimal exponent of n in the number of operations required by any arithmetic algorithm, we thus have $2 \leq \omega \leq 3$. What is the value of ω? Since Strassen published his matrix multiplication algorithm in 1969 we know that $\omega \leq 2.81$ [Str69]. Over the years, more constructions of faster matrix multiplication algorithms, relying on insights involving direct sum algorithms, approximative algorithms and asymptotic induced matchings, lead to the current upper bound $\omega \leq 2.3728639$ [CW90, Sto10, Wil12, LG14].

In applications the matrices to be multiplied are often very rectangular instead of square; see the examples in [LU18]. For any nonnegative real p, given an $n \times \lceil n^p \rceil$ matrix and an $\lceil n^p \rceil \times n$ matrix, how many arithmetic operations are required? Denoting, similarly as in the square case, by $\omega(p)$ the optimal exponent of n in the number of operations required by any arithmetic algorithm, we a priori have the bounds $\max(2, 1+p) \leq \omega(p) \leq 2+p$. (Formally speaking, $\omega(p)$ is the infimum over all real numbers b so that the product of any $n \times \lceil n^p \rceil$ matrix and any $\lceil n^p \rceil \times n$ matrix can be computed in $\mathcal{O}(n^b)$ arithmetic operations. Of course, $\omega = \omega(1)$, and if $\omega = 2$, then $\omega(p) = \max(2, 1+p)$.) What is the value of $\omega(p)$? Parallel to the developments in upper bounding ω, the upper bound $2+p$ was improved drastically over the years for the several regimes of p [HP98, KZHP08, LG12, LU18]. The best lower bound on $\omega(p)$, however, has remained $\max(2, 1+p)$.

ISSN 1433-8092

So the matrix multiplication exponent ω characterises the complexity of square matrix multiplication and, for every nonnegative real p, the rectangular matrix multiplication exponent $\omega(p)$ characterises the complexity of rectangular matrix multiplication. Coppersmith proved that there exists a value $0 < p < 1$ such that $\omega(p) = 2$ [Cop82]. The largest p such that $\omega(p) = 2$ is denoted by α. We will refer to α as the dual matrix multiplication exponent. The algorithms constructed in [LU18] give the currently best bound $\alpha > 0.31389$. If $\alpha = 1$, then of course $\omega = 2$. In fact, $\omega + \frac{\omega}{2}\alpha \leq 3$ (Remark 3.20). Thus we study $\omega(p)$ not only to understand rectangular matrix multiplication, but also as a means to prove $\omega = 2$. The value of α appears explicitly in various applications, for example in the recent work on solving linear programs [CLS19] and empirical risk minimization [LSZ19].

The goal of this paper is to understand why current techniques have not closed the gap between the best lower bound on $\omega(p)$ and the best upper bound on $\omega(p)$, and to thus understand where to find faster rectangular matrix multiplication algorithms. We prove a barrier for current techniques to give much better upper bounds than the current ones. Our work gives a very precise picture of the limitations of current techniques used to obtain the best upper bounds on $\omega(p)$ and the best lower bounds on α.

Our ideas apply as well to $n \times \lceil n^p \rceil$ by $\lceil n^p \rceil \times \lceil n^q \rceil$ matrix multiplication for different p and q. We focus on $p = q$ for simplicity.

1.1. How are algorithms constructed?

To understand what are the current techniques that we prove barriers for, we explain how the fastest algorithms for matrix multiplication are constructed, on a high level. An algorithm for matrix multiplication should be thought of as a reduction of the "matrix multiplication problem" to the natural "unit problem" that corresponds to multiplying numbers,

$$\text{matrix multiplication problem} \leq \text{unit problem}.$$

Mathematically, problems correspond to families of tensors. Several different notions of reduction are used in this context. We will discuss tensors and reductions in more detail later.

In practice, the fastest matrix multiplication algorithms, for square or rectangular matrices, are obtained by a reduction of the matrix multiplication problem to some intermediate problem and a reduction of the intermediate problem to the unit problem,

$$\text{matrix multiplication problem} \leq \text{intermediate problem} \leq \text{unit problem}.$$

The intermediate problems that have been used so far to obtain the best upper bounds on $\omega(p)$ correspond to the so-called small and big Coppersmith–Winograd tensors cw_q and CW_q.

Depending on the intermediate problem and the notion of reduction, we prove a barrier on the best upper bound on $\omega(p)$ that can be obtained in the above way. Before we say something about our new barrier, we discuss the history of barriers for matrix multiplication.

1.2. History of matrix multiplication barriers

We call a lower bound for all upper bounds on ω or $\omega(p)$ that can be obtained by some method, a *barrier* for that method. We give a high-level historical account of barriers for square and rectangular matrix multiplication.

Ambainis, Filmus and Le Gall [AFLG15] were the first to prove a barrier in the context of matrix multiplication. They proved that a variety of methods applied to the Coppersmith–Winograd intermediate tensors (which gave the best upper bounds on ω) cannot give $\omega = 2$ and in fact cannot give $\omega \leq 2.3$.

Alman and Vassilevska Williams [AW18a, AW18b] proved barriers for a notion of reduction called monomial degeneration, extending the realm of barriers beyond the scope of the Ambainis et al. paper. They prove that some collections of intermediate tensors, including the Coppersmith–Winograd intermediate tensors, cannot be used to prove $\omega = 2$. Their analysis is based on studying the so-called asymptotic independence number of the intermediate problem (also called monomial asymptotic subrank). This paper also for the first time studies barriers for *rectangular* matrix multiplication, for $0 \leq p \leq 1$ and monomial degeneration. For example, they prove that the intermediate tensor CW_6 can only give $\alpha \leq 0.87$.

Blasiak et al. [BCC$^+$17a, BCC$^+$17b] did a study of barriers for square matrix multiplication algorithms obtained with a subset of the group-theoretic method, which is a monomial degeneration applied to certain group algebra tensors.

Christandl, Vrana and Zuiddam [CVZ19] proved barriers that apply more generally than the previous one, namely for a type of reduction called degeneration. Their barrier is given in terms of the irreversibility of the intermediate tensor. Irreversibility can be thought of as an asymptotic measure of the failure of Gaussian elimination to bring tensors into diagonal form. To compute irreversibility, they used the asymptotic spectrum of tensors and in particular two families of real tensor parameters with special algebraic properties: the quantum functionals [CVZ18] and support functionals [Str91], although one can equivalently use asymptotic slice rank to compute the barriers for the Coppersmith–Winograd intermediate tensors.

Alman [Alm19] simultaneously and independently obtained the same barrier, relying on a study of asymptotic slice rank.

1.3. New barriers for rectangular matrix multiplication

We prove new barriers for rectangular matrix multiplication using the quantum functionals and support functionals.

We first set up a general barrier framework that encompasses all previously used notions of reductions and then numerically compute barriers for the degeneration notion of reduction and the Coppersmith–Winograd intermediate problems. We also discuss barriers for "mixed" intermediate problems, which covers a method used by, for example, Coppersmith [Cop97].

We will explain our barrier in more detail in the language of tensors, but first we will give a numerical illustration of the barriers.

1.3.1. Numerical illustration of the barriers

For the popular intermediate tensor CW_6 our barrier to get upper bounds on $\omega(p)$ via degeneration looks as follows. In Fig. 1, the horizontal axis goes over all $p \in [0, 2]$. The blue line is the upper bound on $\omega(p)$ obtained via CW_6 as in [LG12]. The yellow line is the barrier and the red line is the best lower bound $\max\{2, 1 + p\}$ on $\omega(p)$. (In [LG12] the best upper bounds on $\omega(p)$ are obtained using CW_q with $q = 5$ for $p \leq 0.81$, $q = 6$ for $0.81 < p \leq 3.5$ and $q = 7$ for $p > 3.5$.)

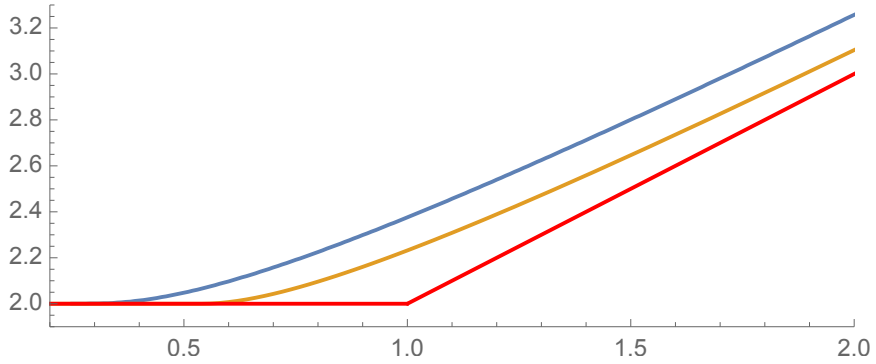

Figure 1. The blue line is the upper bound on $\omega(p)$ obtained via CW_6 as in [LG12] where $p \in [0, 2]$ in on the horizontal axis; the yellow line is our barrier for upper bounds on $\omega(p)$ via degeneration and the intermediate tensor CW_6; the red line is the lower bound on $\omega(p)$.

How about the barrier for CW_q for other values of q? To see what happens there, we give in Fig. 2 the barrier for several values of q in terms of the dual matrix multiplication exponent α. (We recall that α is the largest value of p such that $\omega(p) = 2$.) For $q = 6$ this barrier corresponds to the smallest value of p in Fig. 1 where the yellow line goes above 2.

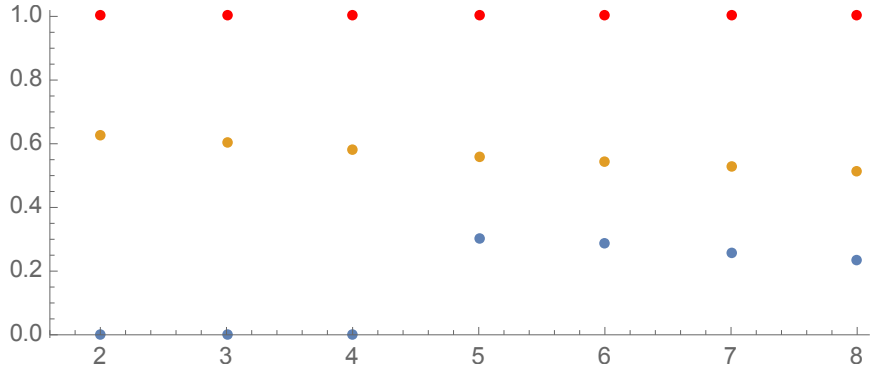

Figure 2. The blue points are the lower bound on α obtained via CW_q as in [LG12] for all $q \in \{2, \ldots, 8\}$, the yellow points are our barrier for the best lower bound on α obtainable via degeneration and the intermediate tensor CW_q, and the red points are the best upper bounds on α, namely 1. The best lower bound $\alpha > 0.3029$ is attained at $q = 5$. Any lower bound on α using degeneration and CW_q for any q, cannot exceed 0.625, the highest yellow point.

Our results give that the best lower bound on α obtainable with degenerations via CW_q for any q, cannot exceed 0.625. (This value corresponds to the highest yellow point in Fig. 2.) Recall that the currently best lower bound is $\alpha > 0.31389$ [LU18].

Compared to [AW18a] our barriers are more general, numerically higher and apply not only for $0 \leq p \leq 1$ but also for $p \geq 1$. For example, [AW18a] proves that monomial degeneration via CW_6 can only give $0.871 \leq \alpha$ whereas we get that the stronger degenerations via CW_6 can only give $0.543 \leq \alpha$.

1.3.2. The barrier in tensor language

Let us continue the discussion that we started in Section 1.1 of how algorithms are constructed, but now in the language of tensors. The goal is to explain our barrier in more detail.

As we mentioned, algorithms correspond to reductions from the matrix multiplication problem to some natural unit problem and the problems correspond to tensors. Let $\mathbb{F}$ be our base field. (The value of $\omega(p)$ may in fact depend on the characteristic of the base field.) A tensor is a trilinear map $\mathbb{F}^{n_1} \times \mathbb{F}^{n_2} \times \mathbb{F}^{n_3} \to \mathbb{F}$. The problem of multiplying an $\ell \times m$ matrix and an $m \times n$ matrix corresponds to the matrix multiplication tensor

$$\langle \ell, m, n \rangle = \sum_{i=1}^{\ell} \sum_{j=1}^{m} \sum_{k=1}^{n} x_{ij} y_{jk} z_{ki}.$$

The unit problem corresponds to the family of diagonal tensors

$$\langle n \rangle = \sum_{i=1}^{n} x_i y_i z_i.$$

There are several notions of reduction that one can consider, but the following is the most natural one. For two tensors S and T we say S is a restriction of T and write $S \leq T$ if there are three linear maps A, B, C of appropriate formats such that S is obtained from T by precomposing with A, B and C, that is, $S = T \circ (A, B, C)$.

A very important observation (see, e.g., [BCS97] or [Blä13]) is that any matrix multiplication algorithm corresponds to an inequality

$$\langle \ell, m, n \rangle \leq \langle r \rangle.$$

Square matrix multiplication algorithms look like

$$\langle n, n, n \rangle \leq \langle r \rangle$$

and rectangular matrix multiplication, of the form that we study, look like

$$\langle n, n, \lceil n^p \rceil \rangle \leq \langle r \rangle.$$

In general, faster algorithms correspond to having smaller r on the right-hand side. In fact, if

$$\langle n, n, n \rangle \leq \langle n^{c+o(1)} \rangle$$

then $\omega \leq c$, and similarly for any $p \geq 0$, if

$$\langle n, n, \lceil n^p \rceil \rangle \leq \langle n^{c+o(1)} \rangle$$

then $\omega(p) \leq c$. For example, if

$$\langle n, n, n^2 \rangle \leq \langle n^{c+o(1)} \rangle$$

then $\omega(2) \leq c$.

Next we utilise a natural product structure on matrix multiplication tensors which is well known as the fact that block matrices can be multiplied block-wise. For tensors S and T one naturally defines a Kronecker product $S \otimes T$ generalizing the matrix Kronecker product. Then the matrix multiplication tensors multiply like $\langle n_1, n_2, n_3 \rangle \otimes \langle m_1, m_2, m_3 \rangle = \langle n_1 m_1, n_2 m_2, n_3 m_3 \rangle$ and the diagonal tensors multiply like $\langle n \rangle \otimes \langle m \rangle = \langle nm \rangle$.

We can thus say: if

$$\langle 2, 2, 2^2 \rangle^{\otimes n} \leq \langle 2 \rangle^{\otimes cn+o(n)}$$

then $\omega(2) \leq c$. We now think of our problem as the problem of determining the optimal asymptotic rate of transformation from $\langle 2 \rangle$ to $\langle 2, 2, 2^2 \rangle$. Of course we can do similarly for values of p other than $p = 2$, if we deal carefully with p that are non-integer. For clarity we will in this section stick to $p = 2$.

In practice, as mentioned before, algorithms are obtained by reductions via intermediate problems. This works as follows. Let T be any tensor, the intermediate tensor. Then clearly, if

$$\langle 2, 2, 2^2 \rangle^{\otimes n} \leq T^{\otimes an+o(n)} \leq \langle 2 \rangle^{\otimes abn+o(n)}, \tag{1}$$

then $\omega(2) \leq ab$. The barrier we prove is a lower bound on ab depending on T and the notion of reduction used in the inequality $\langle 2, 2, 2^2 \rangle^{\otimes n} \leq T^{\otimes an+o(n)}$, which in this section we take to be restriction.

We obtain the barrier as follows. Imagine that F is a map from the set of tensors to the nonnegative real numbers that is $\leq$-monotone, $\otimes$-multiplicative and $\langle n \rangle$-normalised, meaning that for any tensors S and T the following holds: if $S \leq T$ then $F(S) \leq F(T)$; $F(S \otimes T) = F(S)F(T)$ and $F(\langle n \rangle) = n$. We apply F to both sides of the first inequality to get

$$F(\langle 2, 2, 2^2 \rangle) \leq F(T)^a$$

and so

$$\frac{\log F(\langle 2, 2, 2^2 \rangle)}{\log F(T)} \leq a$$

Let G be another map from tensors to reals that is $\leq$-monotone, $\otimes$-multiplicative and $\langle n \rangle$-normalised. We apply G to both sides of the second inequality to get

$$G(T) \leq 2^b$$

and so

$$\log G(T) \leq b.$$

We conclude that

$$\frac{\log F(\langle 2,2,2^2\rangle)}{\log F(T)} \log G(T) \leq ab.$$

Our barrier is thus

$$\max_{F,G} \frac{\log F(\langle 2,2,2^2\rangle)}{\log F(T)} \log G(T) \leq ab.$$

where the maximisation is over the $\leq$-monotone, $\otimes$-multiplicative and $\langle n\rangle$-normalised maps from tensors to reals.

For tensors over the complex numbers, we know a family of $\leq$-monotone, $\otimes$-multiplicative and $\langle n\rangle$-normalised maps from tensors to reals, the quantum functionals. For tensors over other fields, we know a family of maps with slightly weaker properties, that are still sufficient to prove the barrier, the support functionals.

Theorem. *Upper bounds on $\omega(p)$ obtained via the intermediate tensor T are at least*

$$\max_{F,G} \frac{\log(F(\langle 2,1,1\rangle)F(\langle 1,2,1\rangle)F(\langle 1,1,2\rangle)^p)}{\log F(T)} \log G(T)$$

where the maximisation is over all support functionals, or all quantum functionals.

See Theorem 3.13 for the precise statement of the result and Section 1.3.1 for illustrations.

1.3.3. Catalyticity

We discussed that, in practice, the best upper bound on, say, $\omega(2)$ is obtained by a chain of inequalities of the form

$$\langle 2,2,2^2\rangle^{\otimes n} \leq T^{\otimes an+o(n)} \leq \langle 2\rangle^{\otimes abn+o(n)}. \tag{2}$$

We utilised this structure to obtain the barrier. A closer look reveals that the methods used in practice have even more structure. Namely, they give an inequality that also has diagonal tensors on the left-hand side:

$$\langle 2\rangle^{\otimes cn} \otimes \langle 2,2,2^2\rangle^{\otimes n} \leq T^{\otimes an+o(n)} \leq \langle 2\rangle^{\otimes abn+o(n)}. \tag{3}$$

Part of the tensor $\langle 2\rangle^{\otimes abn+o(n)}$ on the far right-hand side acts as a catalyst since $\langle 2\rangle^{\otimes cn}$ is returned on the far left-hand side. We obtain better barriers when we have a handle on the amount of catalyticity c that is used in the method (see the schematic Fig. 3), again by applying maps F and G to both sides of the two inequalities and deducing a lower bound on ab. The precise statement appears in Theorem 3.13.

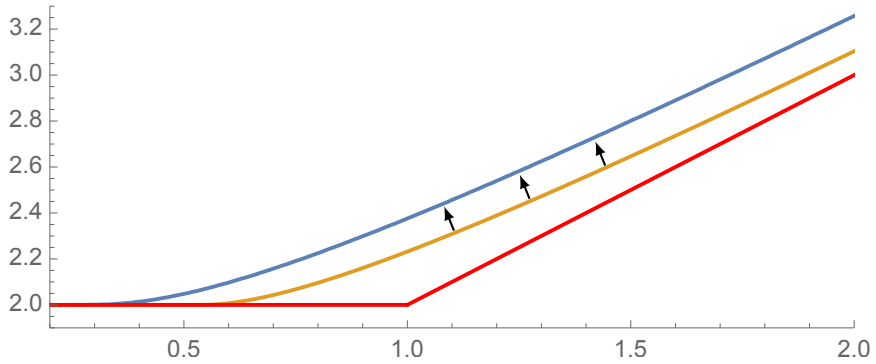

Figure 3. This is the graph from Fig. 1 with arrows that indicate the influence of catalyticity. Roughly speaking, the barrier for CW_6 (the yellow line) moves upwards when more catalyticity is used.

1.4. Overview of the next sections

In Section 2 we discuss in more detail the methods that are used to construct rectangular matrix multiplication algorithms and the different notions of reduction.

In Section 3 we introduce and prove our barriers in the form of a general framework, dealing formally with non-integer p. We also discuss how to analyse "mixed" intermediate tensors.

In Section 4 we discuss how to compute the barriers explicitly using the support functionals and we compute them for the Coppersmith–Winograd tensors CW_q.

2. Algorithms

At the core of the methods that give the best upper bounds on $\omega(p)$ lies the following theorem, which can be proven using the asymptotic sum inequality for rectangular matrix multiplication [LR83] and the monotonicity of $\omega(p)$.

Theorem 2.1. *Let $m \geq n^p$. If $\underset{\sim}{\mathrm{R}}(\langle n,n,m\rangle^{\oplus s}) \leq r$, then $s\, n^{\omega(p)} \leq r$.*

Here $\oplus$ denotes the naturally defined direct sum for tensors. The rank $\mathrm{R}(T)$ of a tensor T is the smallest number n such that $T \leq \langle n\rangle$, or equivalently, the smallest number n such that $T(x,y,z) = \sum_{i=1}^{n} u_i(x)v_i(y)w_i(z)$ where u_i, v_i, w_i are linear. The asymptotic rank $\underset{\sim}{\mathrm{R}}(T)$ is defined as the limit $\lim_{n\to\infty} \mathrm{R}(T^{\otimes n})^{1/n}$, which equals the infimum $\inf_n \mathrm{R}(T^{\otimes n})^{1/n}$ since tensor rank is submultiplicative under $\otimes$ and bounded.

Equivalently, phrased in the language of the introduction, for $m \geq n^p$, if

$$\langle s\rangle^{\otimes k} \otimes \langle n,n,m\rangle^{\otimes k} \leq \langle r\rangle^{\otimes k+o(k)} \tag{4}$$

then $sn^{\omega(p)} < r$. In practice, the upper bound $\underset{\sim}{\mathrm{R}}(\langle n,n,m\rangle^{\oplus s}) \leq r$ is obtained from a restriction $\langle s\rangle^{\otimes k} \otimes \langle n,n,m\rangle^{\otimes k} \leq T^{\otimes ak+o(k)}$ for some intermediate tensor T and an upper bound on $\underset{\sim}{\mathrm{R}}(T)$. The restriction in $\langle s\rangle^{\otimes k} \otimes \langle n,n,m\rangle^{\otimes k} \leq T^{ak+o(k)}$ may be replaced by other types of reductions that we will discuss below.

Reductions. We say S is a *monomial restriction* of T and write $S \leq_M T$ if S can be obtained from T by setting some variables to zero. We say S is a *monomial degeneration* of T and write $S \trianglelefteq_M T$ if S can be obtained from T by multiplying the variables by integer powers of ε so that S appears in the lowest ε-degree. Strassen's application of the laser method uses monomial degenerations and the modification of Coppersmith and Winograd [CW90] uses combinatorial restrictions where the variables zeroed out are chosen using a certain combinatorial gadget (a Salem–Spencer set). *Degeneration* is a very general reduction that generalises the above reductions. We say S is a degeneration of T and write $S \trianglelefteq T$ if S appears in the lowest ε-degree in $T(A(\varepsilon)x, B(\varepsilon)y, C(\varepsilon)z)$ for some linear maps $A(\varepsilon), B(\varepsilon), C(\varepsilon)$ whose matrices have coefficients that are Laurent polynomials in ε. Restriction $\leq$ is the special case of degeneration where the Laurent polynomials are constant.

Coppersmith–Winograd intermediate tensors. All improvements on $\omega(p)$ since Coppersmith and Winograd use the Coppersmith–Winograd tensors CW_q defined by $\mathrm{CW}_q(x, y, z) = x_0y_0z_{q+1} + x_0y_{q+1}z_0 + x_{q+1}y_0z_0 + \sum_{i=1}^{q}(x_0y_iz_i + x_iy_0z_i + x_0y_iz_i)$ as intermediate tensors. Degeneration methods give $\underset{\sim}{\mathrm{R}}(\mathrm{CW}_q) = q + 2$.

Mixed Coppersmith–Winograd tensors. Coppersmith [Cop97] combines CW_q tensors with different q's to upper bound $\omega(p)$. We show how to use the barrier in this situation in the full version. The best upper bounds in [LG12, LU18] do not mix q's.

3. Barriers

Let $\leq$ denote restriction on tensors as defined in the introduction. We remark that everything we discuss in this section also holds if $\leq$ is replaced with degeneration or monomial degeneration or monomial restriction.

Let $F : \{\text{tensors}\} \to \mathbb{R}_{\geq 0}$ be a map from all tensors to the reals. The statements we prove in this section hold for F with certain special properties. Two families that satisfy the properties are the quantum functionals (which we will not explicitly use in this paper — we refer to [CVZ18] for the definition) and the (upper) support functionals. For concreteness, we will think of F as the support functionals. We will define the support functionals in the next section. For now, we will use the following properties.

Lemma 3.1 (Strassen [Str91])**.** *Any support functional F is*

(i) $\leq$*-monotone,*

(ii) $\otimes$*-submultiplicative,*

(iii) *mamu-$\otimes$-multiplicative: F is $\otimes$-multiplicative for any two matrix multiplication tensors,*

(iv) $\oplus$*-additive,*

(v) *at most* $\underset{\sim}{\mathrm{R}}$.

More is known about the support functionals than Lemma 3.1. For example, they are multiplicative not only on the matrix multiplication tensors, but also on a larger family of tensors called oblique tensors.

Remark 3.2. The statements in this section can be proven more generally for certain preorders $\leq$ (including degeneration, monomial degeneration and monomial restriction) and certain maps $F : \{\text{tensors}\} \to \mathbb{R}_{\geq 0}$. Here for concreteness we discuss everything in terms of restriction and the support functionals. A precise discussion will appear in the full version.

3.1. Non-integer p

Recall that p is a nonnegative real number. To deal with p that are not integer we will define a notational shorthand. We first observe the following.

Lemma 3.3. *Let $m \geq n^p$. Suppose that $a \geq 1$ is an integer such that a^p is integer. Then*

$$F(\langle n, n, m\rangle) \geq F(\langle a, a, a^p\rangle)^{\log_a n}.$$

Proof. For every rational number $\frac{s}{t} < \log_a n$ we have

$$F(\langle n, n, m\rangle) = F(\langle n, n, m\rangle^{\otimes t})^{\frac{1}{t}} = F(\langle n^t, n^t, m^t\rangle)^{\frac{1}{t}} \geq F(\langle a^s, a^s, a^{ps}\rangle)^{\frac{1}{t}} = F(\langle a, a, a^p\rangle)^{\frac{s}{t}}. \quad \square$$

From Lemma 3.3 follows that $\log_a F(\langle a, a, a^p\rangle)$ is the same for any a with integer power a^p. We introduce a notation for dealing with this value without referring to the set of possible values of a

Definition 3.4. We introduce a formal symbol $\langle 2, 2, 2^p\rangle$ for each real $p \geq 0$, which we call a quasitensor. If $p = \log_a b$ for integers a and b, then we define

$$F(\langle 2, 2, 2^p\rangle) = 2^{\log_a F(\langle a, a, a^p\rangle)}.$$

Otherwise, we define

$$F(\langle 2, 2, 2^p\rangle) = \inf\{F(\langle 2, 2, 2^P\rangle) : P \geq p,\ P = \log_a b\}.$$

If p is integer, then the value of F on $\langle 2, 2, 2^p\rangle$ as a tensor and as a quasitensor coincide. Thus we identify the quasitensor $\langle 2, 2, 2^p\rangle$ with the tensor $\langle 2, 2, 2^p\rangle$ when the latter exists.

Using this notation, Lemma 3.3 can be rephrased as follows.

Lemma 3.5. *If $m \geq n^p$, then $F(\langle n, n, m\rangle) \geq F(\langle 2, 2, 2^p\rangle)^{\log n}$.*

Lemma 3.6. $F(\langle 2, 2, 2^p\rangle) = F(\langle 2, 1, 1\rangle)F(\langle 1, 2, 1\rangle)F(\langle 1, 1, 2\rangle)^p$.

Proof. We have $F(\langle a,1,1\rangle) = F(\langle 2,1,1\rangle)^{\log a}$ because if $\log a \le \frac{b}{c}$, then $a^c \le 2^b$ and $F(\langle a,1,1\rangle)^c \le F(\langle 2,1,1\rangle)^b$, and if $\log a \ge \frac{b}{c}$, then $F(\langle a,1,1\rangle)^c \ge F(\langle 2,1,1\rangle)^b$. Analogous results hold for $\langle 1,a,1\rangle$ and $\langle 1,1,a\rangle$.

Suppose $p = \log_a b$. Then

$$\begin{aligned}\log F(\langle 2,2,2^p\rangle) = \log_a F(\langle a,a,b\rangle) &= \log_a \left[F(\langle a,1,1\rangle)F(\langle 1,a,1\rangle)F(\langle 1,1,b\rangle)\right] \\ &= \log F(\langle 2,1,1\rangle) + \log F(\langle 1,2,1\rangle) + p\log F(\langle 1,1,2\rangle).\end{aligned}$$

For arbitrary p the result follows by a continuity argument. □

Lemma 3.7. *If $m = n^{p+o(1)}$, then $\log_n F(\langle n,n,m\rangle) = \log F(\langle 2,2,2^p\rangle) + o(1)$.*

Proof. We have

$$F(\langle n,n,m\rangle) = F(\langle n,1,1\rangle)F(\langle 1,n,1\rangle)F(\langle 1,1,m\rangle)$$

and so

$$\begin{aligned}\log_n F(\langle n,n,m\rangle) &= \log F(\langle 2,1,1\rangle) + \log F(\langle 1,2,1\rangle) + \log_n(m)\log F(\langle 1,1,2\rangle) \\ &= \log F(\langle 2,2,2^p\rangle) + o(1)F(\langle 1,1,2\rangle).\end{aligned}$$ □

3.2. T-method

For any tensor T we define the notion of a T-method for upper bounds on $\omega(p)$ as follows.

Definition 3.8 (T-method)**.** Suppose $\underset{\sim}{\mathrm{R}}(T) \le r$. Suppose we are given a collection of inequalities $\langle n,n,m\rangle^{\oplus s} \le T^{\otimes k}$ with $n^p \le m$. Then Theorem 2.1 gives the upper bound $\omega(p) \le \hat{\omega}(p)$ where $\hat{\omega}(p) = \inf\{k\log_n r - \log_n s\}$ where the infimum is taken over all k,n,s appearing in the collection of inequalities. We then say $\hat{\omega}(p)$ is obtained by a *T-method*.

We say that the T-method is *κ-catalytic* if the set of values of n is unbounded, the bound $\hat{\omega}(p)$ is not attained on any one reduction of the method (so $\hat{\omega}(p) = \liminf\{k\log_n r - \log_n s\}$), and in any reduction we have $s \ge Cn^\kappa$ for some constant C.

Theorem 3.9. *Any upper bound $\hat{\omega}(p)$ on $\omega(p)$ obtained by a T-method satisfies*

$$\hat{\omega}(p) \ge \frac{\log F(\langle 2,2,2^p\rangle)\log \underset{\sim}{\mathrm{R}}(T)}{\log F(T)}.$$

Moreover, if the method is κ-catalytic, then

$$\hat{\omega}(p) \ge \frac{\log F(\langle 2,2,2^p\rangle)\log \underset{\sim}{\mathrm{R}}(T)}{\log F(T)} + \kappa\left(\frac{\log \underset{\sim}{\mathrm{R}}(T)}{\log F(T)} - 1\right).$$

Proof. It is enough to prove the inequality for one reduction $T^{\otimes k} \ge \langle n,n,m\rangle^{\oplus s}$ with $m \ge n^p$, which gives an upper bound $\hat{\omega}(p) = k\log_n \underset{\sim}{\mathrm{R}}(T) - \log_n s$.

Using Lemma 3.5 and superadditivity of F, we have

$$F(\langle n,n,m\rangle^{\oplus s}) \geq sF(\langle n,n,m\rangle) \geq sF(\langle 2,2,2^p\rangle)^{\log n}.$$

Therefore $k \log_n F(T) \geq \log_n F(T^{\otimes k}) \geq \log F(\langle 2,2,2^p\rangle) + \log_n s$. For $\hat{\omega}(p)$ we get

$$\frac{\hat{\omega}(p) + \log_n s}{\log F(\langle 2,2,2^p\rangle) + \log_n s} \geq \frac{k \log_n \underset{\sim}{\mathrm{R}}(T)}{k \log_n F(T)} = \frac{\log \underset{\sim}{\mathrm{R}}(T)}{\log F(T)}.$$

Since $F(T) \leq \underset{\sim}{\mathrm{R}}(T)$, we have $\hat{\omega}(p) + \log_n s \geq \log F(\langle 2,2,2^p\rangle) + \log_n s$ and therefore

$$\frac{\hat{\omega}(p)}{\log F(\langle 2,2,2^p\rangle)} \geq \frac{\hat{\omega}(p) + \log_n s}{\log F(\langle 2,2,2^p\rangle) + \log_n s}.$$

If the method is κ-catalytic, then $\log_n s \geq \kappa + O(\frac{1}{\log n})$, and as $n \to \infty$ we have

$$\frac{\hat{\omega}(p) + \kappa}{\log F(\langle 2,2,2^p\rangle) + \kappa} \geq \frac{\log \underset{\sim}{\mathrm{R}}(T)}{\log F(T)}.$$

This concludes the proof. □

3.3. Asymptotic T-method

To cover the method that are used in practice we need the following notion.

Definition 3.10 (Asymptotic T-method.)**.** Let T be a tensor. Suppose $\underset{\sim}{\mathrm{R}}(T) \leq r$. Suppose we are given a collection of inequalities $\langle n,n,m\rangle^{\oplus s} \leq T^{\otimes k}$ where the values of n are unbounded and $m \geq f(n)$ for some function $f(n) = n^{p+o(1)}$. Then $\omega(p)$ is at most $\hat{\omega}(p)$ where $\hat{\omega}(p) = \liminf\{k \log_n r - \log_n s\}$ where the limit is taken over all k, n, s appearing in the collection of inequalities as $n \to \infty$. We say $\hat{\omega}(p)$ is obtained by an *asymptotic T-method*.

We say that the asymptotic T-method is *κ-catalytic* if in any inequality we have $s \geq Cn^{\kappa}$ for some constant C.

Remark 3.11. This class of methods works because each reduction $T^{\otimes k} \geq \langle n,n,m\rangle^{\oplus s}$ gives an upper bound $\omega(q) \leq k \log_n r - \log_n s$ where $q = \log m \geq \log f(n) \to p$. As the function $\omega(p)$ is continuous [LR83], we get the required bound on $\omega(p)$ in the limit.

Remark 3.12. The usual descriptions of the laser method applied to rectangular matrix multiplication result in an asymptotic method because the construction involves an approximation of a certain probability distribution by a rational probability distribution. As a result of this approximation, the matrix multiplication tensor constructed may have format slightly smaller than $\langle n,n,n^p\rangle$.

Theorem 3.13. *Any upper bound $\hat{\omega}(p)$ obtained by an asymptotic T-method satisfies*

$$\hat{\omega}(p) \geq \frac{\log F(\langle 2,2,2^p\rangle) \log \underset{\sim}{\mathrm{R}}(T)}{\log F(T)}$$

and for κ-catalytic methods,

$$\hat{\omega}(p) \geq \frac{\log F(\langle 2,2,2^p\rangle) \log \underset{\sim}{\mathrm{R}}(T)}{\log F(T)} + \kappa \left(\frac{\log \underset{\sim}{\mathrm{R}}(T)}{\log F(T)} - 1 \right).$$

Proof. Suppose $T^k \geq \langle n,n,m\rangle^{\oplus s}$. Then $\hat{\omega}_{k,s,n,m} = k \log_n \underset{\sim}{\mathrm{R}}(T) - \log_n s$ is an upper bound on $\omega(p + o(1))$. Then, as in Theorem 3.9, we have

$$\frac{\hat{\omega}_{k,s,n,m} + \log_n s}{\log_n F(\langle n,n,m\rangle) + \log_n s} \geq \frac{\log \underset{\sim}{\mathrm{R}}(T)}{\log F(T)}.$$

Because $F(T) \leq \underset{\sim}{\mathrm{R}}(T)$, both fractions are greater than 1 and for $0 \leq A \leq \log_n s$ it is true that

$$\frac{\hat{\omega}_{k,s,n,m} + A}{\log_n F(\langle n,n,m\rangle) + A} \geq \frac{\hat{\omega}(p)_{k,s,n,m} + \log_n s}{\log_n F(\langle n,n,m\rangle) + \log_n s}.$$

As $n \to \infty$, we have $\log_n F(\langle n,n,m\rangle) \geq \log F(\langle 2,2,2^p\rangle) + o(1)$ and, if the method is κ-catalytic, then $\log_n s \geq \kappa + o(1)$. The upper bound $\hat{\omega}(p)$ given by the method is the limit $\liminf \hat{\omega}_{k,s,n,m}$. Taking $n \to \infty$, we get the required inequalities. □

3.4. Mixed method

Coppersmith [Cop97] uses a combination of Coppersmith–Winograd tensors of different format to get an upper bound on the rectangular matrix multiplication exponent. More specifically, he considers a sequence of tensors $\mathrm{CW}_7^{\otimes 9n} \otimes \mathrm{CW}_6^{\otimes 8\lfloor 0.6425n\rfloor}$. Our analysis applies to tensor sequences of this kind because their asymptotic behaviour is similar to sequence of the form $T^{\otimes n}$ in the sense of the following two lemmas.

Lemma 3.14. *Let S_1 and S_2 be some tensors. Given functions $f_1, f_2 \colon \mathbb{N} \to \mathbb{N}$ such that $f_i(n) = a_i n + o(n)$ for some positive real numbers a_1, a_2, define a sequence of tensors $T_n = S_1^{\otimes f_1(n)} \otimes S_2^{\otimes f_2(n)}$. Then for each F the sequence $\sqrt[n]{F(T_n)}$ is bounded from above.*

Proof. We have

$$\sqrt[n]{F(T_n)} = \sqrt[n]{F(S_1^{\otimes f_1(n)} \otimes S_2^{\otimes f_2(n)})} \leq F(S_1)^{\frac{f_1(n)}{n}} F(S_2)^{\frac{f_2(n)}{n}}.$$

The right-hand side converges to $F(S_1)^{a_2} F(S_2)^{a_2}$ and, therefore, is bounded. □

Lemma 3.15. *Let S_1 and S_2 be some tensors. Given functions $f_1, f_2 \colon \mathbb{N} \to \mathbb{N}$ such that $f_i(n) = a_i n + o(n)$ for some positive real numbers a_1, a_2, define a sequence of tensors $T_n = S_1^{\otimes f_1(n)} \otimes S_2^{\otimes f_2(n)}$. Then the sequence $\sqrt[n]{\underset{\sim}{\mathrm{R}}(T_n)}$ converges.*

Proof. For this, we need Strassen's spectral characterization of the asymptotic rank [Str88]. Strassen defines the asymptotic spectrum of tensors X as the set of all $\leq$-monotone, $\otimes$-multiplicative, $\oplus$-additive maps ξ from tensors to positive reals such that $\xi(u \otimes v \otimes w) = 1$. Then X can be made into a compact Hausdorff

topological space such that the evaluation map $\xi \mapsto \xi(T)$ is continuous for all T, and

$$\underset{\sim}{\mathrm{R}}(T) = \max_{\xi \in X} \xi(T),$$

For $\xi \in X$ we have

$$\sqrt[n]{\xi(T_n)} = \sqrt[n]{\xi(S_1^{\otimes f_1(n)} \otimes S_2^{\otimes f_2(n)})} = \xi(S_1)^{\frac{f_1(n)}{n}} \xi(S_2)^{\frac{f_2(n)}{n}} \to \xi(S_1)^{a_1} \xi(S_2)^{a_2}.$$

Because of compactness of X this convergence is uniform in ξ. Therefore

$$\sqrt[n]{\underset{\sim}{\mathrm{R}}(T_n)} = \sqrt[n]{\max_{\xi \in X} \xi(T_n)} \to \max_{\xi \in X} \xi(S_1)^{a_1} \xi(S_2)^{a_2}. \qquad \square$$

Definition 3.16. A sequence of tensors $\{T_n\}$ is called *almost exponential* if the sequences $\sqrt[n]{\underset{\sim}{\mathrm{R}}(T_n)}$ converges and $\sqrt[n]{F(T_n)}$ is bounded for each F. Abusing the notation, we write $\underset{\sim}{\mathrm{R}}(\{T_n\}) \coloneqq \lim \sqrt[n]{\underset{\sim}{\mathrm{R}}(T_n)}$ and $F(\{T_n\}) \coloneqq \limsup \sqrt[n]{F(T_n)}$.

Definition 3.17 (Asymptotic mixed method)**.** Let $\{T_n\}$ be an almost exponential sequence of tensors with $\underset{\sim}{\mathrm{R}}(\{T_n\}) \leq r$. Suppose we are given a collection of inequalities $\langle n, n, m \rangle^{\oplus s} \leq T_k$ where the values of n are unbounded and $m \geq f(n)$ for some $f(n) = n^{p+o(1)}$. Then $\omega(p)$ is at most $\hat{\omega}(p) = \liminf\{k \log_n r - \log_n s\}$ where the limit is taken over all k, n, s appearing in the collection of inequalities as $n \to \infty$. We say that $\hat{\omega}(p)$ is obtained by an *asymptotic mixed* $\{T_n\}$*-method*.

We say that the asymptotic mixed $\{T_n\}$-method is κ*-catalytic* if in each inequality we have $s \geq Cn^{\kappa}$ for some constant C.

Lemma 3.18. *Asymptotic mixed methods give true upper bounds on* $\omega(p)$*.*

Proof. Note that for a fixed tensor T_k there are only a finite number of restrictions $\langle n, n, m \rangle^{\oplus s} \leq T_k$ possible as the left tensor is of format $sn^2 \times snm \times snm$, which should be no greater than the format of T_k. Thus, because in an asymptotic mixed method the set of values of n is unbounded, so is the set of values of k.

For one restriction $\langle n, n, m \rangle^{\oplus s} \leq T_k$ we have the inequality $sn^{\omega(\log_n m)} \leq \underset{\sim}{\mathrm{R}}(T_k)$, that is, $\omega(\log_n m) \leq \log_n \underset{\sim}{\mathrm{R}}(T_k) - \log_n s$. Since $\log_n m = p + o(1)$ and ω is a continuous function and $\underset{\sim}{\mathrm{R}}(T_k) = (\underset{\sim}{\mathrm{R}}(\{T_k\}) + o(1))^k$, we get in the limit the required inequality. $\square$

Theorem 3.19. *Any upper bound* $\hat{\omega}(p)$ *obtained by an asymptotic mixed* $\{T_n\}$*-method satisfies*

$$\hat{\omega}(p) \geq \frac{\log F(\langle 2, 2, 2^p \rangle) \log \underset{\sim}{\mathrm{R}}(\{T_n\})}{\log F(\{T_n\})}$$

and for κ*-catalytic methods,*

$$\hat{\omega}(p) \geq \frac{\log F(\langle 2, 2, 2^p \rangle) \log \underset{\sim}{\mathrm{R}}(\{T_n\})}{\log F(\{T_n\})} + \kappa \left(\frac{\log \underset{\sim}{\mathrm{R}}(\{T_n\})}{\log F(\{T_n\})} - 1 \right).$$

Proof. Recall that for a fixed T_k the number of possible restrictions $\langle n,n,m\rangle^{\oplus s} \le T_k$ is finite, as the left-hand side tensor has format $sn^2 \times snm \times snm$, which should be no greater than that of T_k. Therefore, as n tends to infinity, so does k.

Consider now one restriction $\langle n,n,m\rangle^{\oplus s} \le T_k$. It gives the upper bound $\hat{\omega}_{k,s,n,m} \coloneqq \log_n \underset{\sim}{\mathrm{R}}(T_k) - \log_n s$ on $\omega(p+o(1))$. As in previous theorems, we have

$$\frac{\hat{\omega}_{k,s,n,m} + \log_n s}{\log_n F(\langle n,n,m\rangle) + \log_n s} \ge \frac{\log \underset{\sim}{\mathrm{R}}(T_k)}{\log F(T_k)}$$

and

$$\frac{\hat{\omega}_{k,s,n,m} + A}{\log_n F(\langle n,n,m\rangle) + A} \ge \frac{\hat{\omega}(p)_{k,s,n,m} + \log_n s}{\log_n F(\langle n,n,m\rangle) + \log_n s}$$

for any A such that $0 \le A \le \log_n s$.

Consider the behaviour of the involved quantities as n and k tend to infinity. Since $m \ge n^{p+o(1)}$, $\log_n F(\langle n,n,m\rangle) \ge \log F(\langle 2,2,2^p\rangle) + o(1)$. For a catalytic method, we can choose $A = \kappa + o(1)$ such that $\log_n s \ge A$, and in general, we set $A = 0$. Since $\sqrt[k]{\underset{\sim}{\mathrm{R}}(T_k)} = \underset{\sim}{\mathrm{R}}(\{T_k\}) + o(1)$ and $\sqrt[k]{F(T_k)} \le F(\{T_k\}) + o(1)$, we have

$$\frac{\log \underset{\sim}{\mathrm{R}}(T_k)}{\log F(T_k)} \ge \frac{\log \underset{\sim}{\mathrm{R}}(\{T_k\})}{\log F(\{T_k\})} + o(1).$$

And finally, $\liminf \hat{\omega}_{k,s,n,m}$ is $\hat{\omega}(p)$. In the limit, we get the required inequalities. □

3.5. Barriers on α

The barriers for the lower bounds on the dual matrix multiplication exponent α follow from the barriers for upper bounds on $\omega(p)$. A method can prove the lower bound $\alpha \ge \hat{\alpha}$ on α if it can prove $\omega(\hat{\alpha}) = 2$. For our barrier this means that

$$\frac{\log F(\langle 2,2,2^{\hat{\alpha}}\rangle) \log \underset{\sim}{\mathrm{R}}(T)}{\log F(T)} \le 2$$

for all F. Using Lemma 3.6, we get that any lower bound $\hat{\alpha}$ obtained by an asymptotic T-method satisfies

$$\hat{\alpha} \le \frac{2\log F(T)}{\log \underset{\sim}{\mathrm{R}}(T) \log F(\langle 1,1,2\rangle)} - \frac{\log F(\langle 2,2,1\rangle)}{\log F(\langle 1,1,2\rangle)}$$

for all F such that $\log F(\langle 1,1,2\rangle) \ne 0$.

Remark 3.20. We note in passing that the matrix multiplication exponent ω and the dual exponent α are related via the inequality $\omega + \frac{\omega}{2}\alpha \le 3$. Namely, from $\langle \lceil n^\alpha \rceil, n, n\rangle \le \langle n^{2+o(1)}\rangle$, $\langle n, \lceil n^\alpha \rceil, n\rangle \le \langle n^{2+o(1)}\rangle$ and $\langle n, n, \lceil n^\alpha \rceil\rangle \le \langle n^{2+o(1)}\rangle$ it follows that $\langle n^{2+\alpha}, n^{2+\alpha}, n^{2+\alpha}\rangle \le \langle n^{6+o(1)}\rangle$. Therefore, $\omega \le 6/(2+\alpha)$, and the claim follows.

4. Numerical computation of barriers

We will in this section show how to numerically evaluate the barrier of Theorem 3.13. We will compute explicit values for the Coppersmith–Winograd tensors.

4.1. Upper support functionals

Our main tool is a family of maps called the upper support functionals, introduced by Strassen in [Str91]. To define them, we will use the following notation. For $n \in \mathbb{N}$ let $[n] \coloneqq \{1, 2, \ldots, n\}$. For any finite set A let $\mathcal{P}(A)$ be the set of probability vectors on A. For finite sets A_1, A_2, A_3 and $P \in \mathcal{P}(A_1 \times A_2 \times A_3)$ let $P_i \in \mathcal{P}(A_i)$ be the ith marginal of P for $i \in [3]$. Let $H(P)$ denote the Shannon entropy of P.

Let $\mathbb{F}^{n\times n\times n}$ be the set of 3-tensors of dimension $n\times n\times n$, viewed as 3-dimensional arrays. For $T \in \mathbb{F}^{n\times n\times n}$ let $\operatorname{supp}(T) \subseteq [n]^3$ be the support of T.

Let $T \in \mathbb{F}^{n\times n\times n}$. Let $\theta = (\theta_1, \theta_2, \theta_3) \in \mathcal{P}([3])$. Define

$$\zeta^\theta(T) = \min_{S \cong T} \max_{P \in \mathcal{P}(\operatorname{supp}(S))} 2^{\sum_{i\in[3]} \theta_i H(P_i)} \tag{5}$$

where S goes over all tensors that can be obtained from T by a basis transformation, that is, $S = T \circ (A, B, C)$ where A, B, C are invertible linear maps. The map ζ^θ is called the *upper support functional*.

4.2. Rectangular matrix multiplication

Lemma 4.1. $\zeta^\theta(\langle a, b, c\rangle) = a^{\theta_1+\theta_3} b^{\theta_1+\theta_2} c^{\theta_2+\theta_3}$

Proof. One verifies this by a direct computation. See also [Str91]. □

We obtain from Theorem 3.13 and Lemma 3.6 that any upper bound $\hat{\omega}(p)$ on $\omega(p)$ obtained by asymptotic T-methods must satisfy

$$\hat{\omega}(p) \geq \frac{\log \zeta^\theta(\langle 2, 2, 2^p\rangle)}{\log \zeta^\theta(T)} \log \underset{\sim}{\mathrm{R}}(T),$$

which gives

$$\hat{\omega}(p) \geq \max_\theta \frac{2\theta_1 + \theta_3 + \theta_2 + p(\theta_2 + \theta_3)}{\log_2 \zeta^\theta(T)} \log_2 \underset{\sim}{\mathrm{R}}(T). \tag{6}$$

4.3. Symmetry in the convex program

Before we talk about computations for CW_q we briefly discuss the standard way to make use of symmetry in the optimisation problems that we need to solve. We will be interested in computing

$$\max_{P \in \mathcal{P}(\operatorname{supp}(\mathrm{CW}_q))} \sum_{i=1}^{3} \theta_i H(P_i). \tag{7}$$

Recall that the support of CW_q is

$$\mathrm{supp}(\mathrm{CW}_q) = \{(i,i,0),(i,0,i),(0,i,i) : i \in [q]\} \cup \{(0,0,q+1),(0,q+1,0),(q+1,0,0)\}.$$

The symmetric group S_q acts naturally on the support of CW_q by permuting the label set $[q]$. Suppose P is feasible for (7). Then $\pi \cdot P$ for any $\pi \in S_q$ is feasible as well and has the same value. Thus

$$\frac{1}{|S_q|} \sum_{\pi \in S_q} \pi \cdot P \tag{8}$$

is feasible and has at least the same value or better, by concavity of H. We may thus assume that P is constant on the six orbits of $\mathrm{supp}(\mathrm{CW}_q)$ under the action of S_q, which are the sets $\{(i,i,0) : i \in [q]\}$, $\{(i,0,i) : i \in [q]\}$, $\{(0,i,i) : i \in [q]\}$, $\{(0,0,q+1)\}$, $\{(0,q+1,0)\}$, and $\{(q+1,0,0)\}$. The same reasoning applies when CW_q is replaced by any tensor with symmetry.

4.4. Barriers for CW_q

Taking into account the symmetry derived in Section 4.3, let P be the probability distribution that gives probability p_1 to $(0,i,i)$, probability p_2 to $(i,0,i)$, probability p_3 to $(i,i,0)$ and probability r_1 to $(q+1,0,0)$, probability r_2 to $(0,q+1,0)$ and probability r_3 to $(0,0,q+1)$ where $p_1,p_2,p_3,r_1,r_2,r_3 \geq 0$ and $qp_1+qp_2+qp_3+r_1+r_2+r_3 = 1$. The marginal probability vectors are

$$P_1 = (qp_1 + r_2 + r_3, p_2 + p_3, \ldots, p_2 + p_3, r_1) \tag{9}$$
$$P_2 = (qp_2 + r_1 + r_3, p_1 + p_3, \ldots, p_1 + p_3, r_2) \tag{10}$$
$$P_3 = (qp_3 + r_1 + r_2, p_1 + p_2, \ldots, p_1 + p_2, r_3). \tag{11}$$

By the grouping property of Shannon entropy, we have

$$H(P_1) = (1 - qp_1 - r_2 - r_3)(\log_2(q) + h(r_1)) + h(qp_1 + r_2 + r_3) \tag{12}$$
$$H(P_2) = (1 - qp_2 - r_1 - r_3)(\log_2(q) + h(r_2)) + h(qp_2 + r_1 + r_3) \tag{13}$$
$$H(P_3) = (1 - qp_3 - r_1 - r_2)(\log_2(q) + h(r_3)) + h(qp_3 + r_1 + r_2) \tag{14}$$

and

$$\log_2 \zeta^\theta(\mathrm{CW}_q) \leq \max_{p_j,r_j} \sum_{i=1}^{3} \theta_i H(P_i) \tag{15}$$

where $p_1,p_2,p_3,r_1,r_2,r_3 \geq 0$ and $qp_1 + qp_2 + qp_3 + r_1 + r_2 + r_3 = 1$. We know that $\underset{\sim}{\mathrm{R}}(\mathrm{CW}_q) = q+2$.

The barrier we get for CW_q is

$$\hat{\omega}(p) \geq \max_\theta \frac{2\theta_1 + (p+1)(\theta_2+\theta_3)}{\log_2 \zeta^\theta(\mathrm{CW}_q)} \log_2 \underset{\sim}{\mathrm{R}}(\mathrm{CW}_q) \tag{16}$$
$$\geq \max_\theta \frac{2\theta_1 + (p+1)(\theta_2+\theta_3)}{\max_{p_j,r_j} \sum_{i=1}^{3} \theta_i H(P_i)} \log_2(q+2), \tag{17}$$

which is easy to evaluate numerically.

As an illustration, we give in Table 1 the barriers for upper bounds on $\omega(2)$ via asymptotic CW_q-methods for small q by numerical optimisation. Optimal values were obtained for θ with $\theta_2 = \theta_3$.

q	barrier	θ_1	$\theta_2 = \theta_3$
1	3.0551	0.09	0.455
2	3.0625	0.1	0.45
3	3.0725	0.11	0.445
4	3.0831	0.12	0.44
5	3.0936	0.13	0.435
6	3.1038	0.14	0.43
7	3.1137	0.14	0.43
8	3.1232	0.15	0.425
9	3.1322	0.16	0.42
10	3.1408	0.17	0.415
11	3.1491	0.17	0.415
12	3.1568	0.18	0.41
13	3.1643	0.18	0.41
14	3.1713	0.18	0.41

Table 1. Barriers for upper bounds on $\omega(2)$ via asymptotic CW_q-methods for small q.

Acknowledgements

MC and VL were supported by VILLUM FONDEN via the QMATH Centre of Excellence under Grant No. 10059 and the European Research Council (Grant agreement No. 818761).

FLG was supported by JSPS KAKENHI grants Nos. JP15H01677, JP16H01705, JP16H05853, JP19H04066 and by the MEXT Quantum Leap Flagship Program (MEXT Q-LEAP) grant No. JPMXS0118067394.

JZ was supported by National Science Foundation under Grant No. DMS-1638352.

References

[AFLG15] Andris Ambainis, Yuval Filmus, and François Le Gall. Fast matrix multiplication: limitations of the Coppersmith-Winograd method (extended abstract). In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC 2015)*, pages 585–593, 2015. `arXiv:1411.5414`.

[Alm19] Josh Alman. Limits on the Universal Method for Matrix Multiplication. In *Proceedings of the 34th Computational Complexity Conference (CCC 2019)*, pages 12:1–12:24, 2019. `arXiv:1812.08731`.

[AW18a] Josh Alman and Virginia Vassilevska Williams. Further Limitations of the Known Approaches for Matrix Multiplication. In *Proceedings of the 9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, pages 25:1–25:15, 2018. arXiv:1712.07246.

[AW18b] Josh Alman and Virginia Vassilevska Williams. Limits on All Known (and Some Unknown) Approaches to Matrix Multiplication. In *Proceedings of the 59th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2018)*, pages 580–591, 2018. arXiv:1810.08671.

[BCC+17a] Jonah Blasiak, Thomas Church, Henry Cohn, Joshua A. Grochow, Eric Naslund, William F. Sawin, and Chris Umans. On cap sets and the group-theoretic approach to matrix multiplication. *Discrete Anal.*, 2017. arXiv:1605.06702.

[BCC+17b] Jonah Blasiak, Thomas Church, Henry Cohn, Joshua A Grochow, and Chris Umans. Which groups are amenable to proving exponent two for matrix multiplication? *arXiv*, 2017. arXiv:1712.02302.

[BCS97] Peter Bürgisser, Michael Clausen, and M. Amin Shokrollahi. *Algebraic complexity theory*, volume 315 of *Grundlehren Math. Wiss.* Springer-Verlag, Berlin, 1997.

[Blä13] Markus Bläser. *Fast Matrix Multiplication*. Number 5 in Graduate Surveys. Theory of Computing Library, 2013.

[CLS19] Michael B. Cohen, Yin Tat Lee, and Zhao Song. Solving Linear Programs in the Current Matrix Multiplication Time. In *Proceedings of the 51st Annual ACM Symposium on Theory of Computing (STOC 2019)*, page 938–942, 2019. arXiv:1810.07896.

[Cop82] Don Coppersmith. Rapid Multiplication of Rectangular Matrices. *SIAM J. Comput.*, 11(3):467–471, 1982.

[Cop97] Don Coppersmith. Rectangular Matrix Multiplication Revisited. *J. Complexity*, 13(1):42–49, 1997.

[CVZ18] Matthias Christandl, Péter Vrana, and Jeroen Zuiddam. Universal points in the asymptotic spectrum of tensors. In *Proceedings of the 50th Annual ACM Symposium on Theory of Computing (STOC 2018)*, pages 289–296, 2018. arXiv:1709.07851.

[CVZ19] Matthias Christandl, Péter Vrana, and Jeroen Zuiddam. Barriers for Fast Matrix Multiplication from Irreversibility. In *Proceedings of the 34th Computational Complexity Conference (CCC 2019)*, pages 26:1–26:17, 2019. arXiv:1812.06952.

[CW90] Don Coppersmith and Shmuel Winograd. Matrix Multiplication via Arithmetic Progressions. *J. Symb. Comput.*, 9(3):251–280, 1990.

[HP98] Xiaohan Huang and Victor Y. Pan. Fast Rectangular Matrix Multiplication and Applications. *J. Complexity*, 14(2):257–299, 1998.

[KZHP08] ShanXue Ke, BenSheng Zeng, WenBao Han, and Victor Y. Pan. Fast rectangular matrix multiplication and some applications. *Science in China Series A: Mathematics*, 51(3):389–406, 2008.

[LG12] François Le Gall. Faster algorithms for rectangular matrix multiplication. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS 2012)*, pages 514–523, 2012. `arXiv:1204.1111`.

[LG14] François Le Gall. Powers of tensors and fast matrix multiplication. In *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation (ISSAC 2014)*, pages 296–303, 2014. `arXiv:1401.7714`.

[LR83] Grazia Lotti and Francesco Romani. On the Asymptotic Complexity of Rectangular Matrix Multiplication. *Theor. Comput. Sci.*, 23:171–185, 1983.

[LSZ19] Yin Tat Lee, Zhao Song, and Qiuyi Zhang. Solving Empirical Risk Minimization in the Current Matrix Multiplication Time. In *Conference on Learning Theory (COLT 2019)*, pages 2140–2157, 2019. `arXiv:1905.04447`.

[LU18] François Le Gall and Florent Urrutia. Improved Rectangular Matrix Multiplication using Powers of the Coppersmith-Winograd Tensor. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2018)*, pages 1029–1046, 2018. `arXiv:1708.05622`.

[Sto10] Andrew James Stothers. *On the complexity of matrix multiplication.* PhD thesis, University of Edinburgh, 2010. `http://hdl.handle.net/1842/4734`.

[Str69] Volker Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 13(4):354–356, 1969.

[Str88] Volker Strassen. The asymptotic spectrum of tensors. *J. reine angew. Math.*, 384:102–152, 1988.

[Str91] Volker Strassen. Degeneration and complexity of bilinear maps: some asymptotic spectra. *J. Reine Angew. Math.*, 413:127–180, 1991.

[Wil12] Virginia Vassilevska Williams. Multiplying matrices faster than Coppersmith-Winograd (extended abstract). In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC 2012)*, pages 887–898, 2012.

Matthias Christandl
University of Copenhagen,
Universitetsparken 5, 2100 Copenhagen Ø, Denmark
Email: christandl@math.ku.dk

François Le Gall
Nagoya University,
Furocho, Chikusaku, Nagoya Aichi 464-8602, Japan
Email: legall@math.nagoya-u.ac.jp

Vladimir Lysikov
University of Copenhagen,
Universitetsparken 5, 2100 Copenhagen Ø, Denmark
Email: vl@math.ku.dk

Jeroen Zuiddam
Institute for Advanced Study,
1 Einstein Drive, Princeton, NJ 08540, USA
Email: jzuiddam@ias.edu

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il