On the Coppersmith–Winograd method

Andris Ambainis, University of Latvia* Yuval Filmus, Institute for Advanced Study

May 29, 2014

Abstract

Until a few years ago, the fastest known matrix multiplication algorithm, due to Coppersmith and Winograd (1990), ran in time $O(n^{2.376})$. Recently, a surge of activity by Stothers, Vassilevska-Williams and Le Gall has led to an improved algorithm running in time $O(n^{2.3728639})$, due to Le Gall (2014). These algorithms are obtained by analyzing higher and higher tensor powers of a certain identity of Coppersmith and Winograd. We show that this approach cannot result in an algorithm with running time $O(n^{2.3078})$, and in particular cannot prove the conjecture that for every $\epsilon > 0$, matrices can be multiplied in time $O(n^{2+\epsilon})$.

We describe a new framework extending the original laser method, which is the method underlying the previously mentioned algorithms. Our framework accommodates the algorithms by Coppersmith and Winograd, Stothers, Vassilevska-Williams and Le Gall. We obtain our main result by analyzing this framework. The framework is also the first to explain why taking tensor powers of the Coppersmith–Winograd identity results in faster algorithms.

1 Introduction

How fast can we multiply two $n \times n$ matrices? Ever since Strassen [10] improved on the $O(n^3)$ high-school algorithm, this question has captured the imagination of computer scientists. A theory of fast algorithms for matrix multiplication has been developed. Highlights include Schönhage's asymptotic sum inequality [8], Strassen's laser method [12], and the Coppersmith–Winograd algorithm [4]. The algorithm by Coppersmith and Winograd had been the world champion for 20 years, until finally being improved by Stothers [9] in 2010. Independently, Vassilevska-Williams [13] obtained a further improvement in 2012, and Le Gall [7] perfected their methods to obtain the current world champion in 2014.

The Coppersmith–Winograd algorithm relies on a certain identity which we call the *Coppersmith–Winograd identity*. Using a very clever combinatorial construction and the laser method technique, Coppersmith and Winograd were able to extract a fast matrix multiplication algorithm whose running time is $O(n^{2.388})$. Applying their technique recursively for the tensor square of their identity, they obtain an even faster matrix multiplication algorithm with running time $O(n^{2.376})$. For a long time, this latter algorithm had been the state of the art. The calculations for higher tensor powers are complicated, and yield no improvement for the tensor cube. With the advent of modern computers, however, it became possible to automate the necessary calculations, allowing Stothers to analyze the fourth tensor power and obtain an algorithm with running time $O(n^{2.37293})$. Apart from implementing the necessary computer programs, Stothers also had to generalize the original framework of Coppersmith and Winograd. Independently, Vassilevska-Williams performed the necessary calculations for the fourth and eighth tensor powers, obtaining an algorithm with running time $O(n^{2.3728642})$ for the latter. Higher tensor powers require more extensive calculations, involving the approximate solution of large optimization problems. Le Gall came up with a faster method for solving these large optimization problems (albeit yielding slightly worse solutions), and this enabled him to perform the necessary calculations for the sixteenth and thirty-second tensor powers, obtaining algorithms with running times $O(n^{2.3728640})$ and $O(n^{2.3728639})$, respectively.

*Research done while visiting the Institute for Advanced Study.

It is commonly conjectured that for every $\epsilon > 0$, there exists a matrix multiplication algorithm with running time $O(n^{2+\epsilon})$. Can taking higher and higher tensor powers yield these algorithms? In this paper we answer this question in the negative. We show that taking the Nth tensor power cannot yield an algorithm with running time $O(n^{2.3078})$, for *any* value of N. We obtain this lower bound by presenting a framework which generalizes the techniques of Coppersmith and Winograd, Stothers, Vassilevska-Williams and Le Gall, and is amenable to analysis. At the same time, our framework is the first to explain what is gained by taking tensor powers of the original Coppersmith–Winograd identity.

The Coppersmith–Winograd identity bounds the *border rank* (a certain measure of complexity) of a certain *tensor* (three-dimensional analog of a matrix) T_{CW}. The tensor is a sum of six non-disjoint smaller tensors. Schönhage's asymptotic sum inequality allows us to obtain a matrix multiplication algorithm given a bound on the border rank of a sum of *disjoint* tensors of a special kind, which includes the tensors appearing in T_{CW}. The idea of the laser method is to take a high tensor power of T_{CW} and zero out some of the variables so that the surviving smaller tensors are disjoint. Applying Schönhage's asymptotic sum inequality then yields a matrix multiplication algorithm. Following this route, we obtain an algorithm with running time $O(n^{2.388})$.

In order to improve on this, Coppersmith and Winograd take the tensor square of T_{CW}, and rewrite it as a sum of fifteen non-disjoint smaller tensors, which result from merging in a particular way the thirty-six tensors obtained from the squaring. At this point we repeat the earlier construction. In total, the new construction is equivalent to the following procedure. Starting with the original tensor T_{CW}, we take a high tensor power, zero out some of the variables, and merge groups of remaining tensors so that the resulting merged tensors are disjoint and are of the kind that allows application of the asymptotic sum inequality. We can call this method the *laser method with merging*. The further constructions of Stothers, Vassilevska-Williams and Le Gall can all be put in this framework.

Cohn, Kleinberg, Szegedy and Umans [3] analyzed the simple construction of Coppersmith and Winograd (corresponding to the $O(n^{2.388})$ algorithm), showing that their construction is optimal in the framework of the laser method. Using similar but more complicated ideas, we are able to give an analogous bound for the laser method with merging, showing that the method cannot yield an algorithm with running time $O(n^{2.3078})$. Unfortunately, we are not able to come up with an algorithm matching our barrier, but we believe that our new framework could lead to improved algorithms in the future.

The Coppersmith–Winograd identity is parameterized by an integer parameter $q \geq 0$. Our method applies for all values $q \geq 2$. The bound $O(n^{2.3078})$ corresponds to the choice $q = 5$, which is the value used by Coppersmith and Winograd, Stothers, Vassilevska-Williams and Le Gall. Our bound deteriorates as q gets smaller, and for $q = 2$ we only obtain a limit of $O(n^{2.254})$. We do not know whether this deterioration results from the fact that our bound is not tight, or whether smaller values of q actually lead to better algorithms.

Paper organization Section 2 describes the theory of fast matrix multiplication up to and including the simple construction of Coppersmith and Winograd. The laser method with merging is described in Section 3, in which we also explain how the algorithms of Coppersmith and Winograd, Stothers, Vassilevska-Williams and Le Gall fit in this framework. We prove our main result in Section 4. We close the paper in Section 5 by discussing our results and their implication.

Acknowledgements This material is based upon work supported by the National Science Foundation under agreement No. DMS-1128155. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Science Foundation.

We thank François Le Gall, Edinah Gnang and Avi Wigderson for helpful discussions.

2 Background

Notation We define $[n] = \{1, \ldots, n\}$. We will use the notation $\exp_2 x$ for 2^x. The entropy function H is given by

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i,$$

where $0 \log 0 = 0$. It can be used to estimate multinomial coefficients:

$$\binom{n}{np_1, \ldots np_m} \leq \exp_2 H(p_1, \ldots, p_m)n.$$

The entropy function is *concave*: if $\vec{q}_1, \ldots, \vec{q}_r$ are probability distributions and $c_1, \ldots, c_r \geq 0$ sum to 1 then

$$\sum_{i=1}^{r} c_i H(\vec{q}_i) \leq H\left(\sum_{i=1}^{r} c_i \vec{q}_i\right).$$

2.1 Bilinear complexity

The material below can be found in Chapters 14–15 of the book Algebraic Complexity Theory [2].

The model Our goal in this paper is to study the complexity of matrix multiplication in the algebraic complexity model. In this model, a program for computing the product $C = AB$ of two $n \times n$ matrices is allowed to use the following instructions:

- Reading the input: $t \leftarrow a_{ij}$ or $t \leftarrow b_{ij}$.
- Arithmetic: $t \leftarrow t_1 \circ t_2$, where $\circ \in \{+, -, \times, \div\}$.
- Output: $c_{ij} \leftarrow t$.

Each of these instructions has unit cost. A legal program is one which never divides by zero; Strassen [11] showed how to eliminate divisions at the cost of a constant blowup in the size. Denote by $T(n)$ the size of the smallest program which computes the product of two $n \times n$ matrices. The *exponent of matrix multiplication* is defined by

$$\omega = \lim_{n \to \infty} T(n)^{1/n}.$$

It can be shown that the limit indeed exists. For each $\epsilon > 0$, we also have $T(n) = O_\epsilon(n^{\omega+\epsilon})$, and ω can also be defined via this property.

Tensors Strassen [10] related ω to the tensor rank of *matrix multiplication tensors*, a connection we proceed to explain. The tensors we are interested in are three-dimensional equivalents of matrices. An $n \times m$ matrix A corresponds to the bilinear form

$$x'Ay = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} x_i y_j.$$

Similarly, third order tensors correspond to $n \times m \times p$ trilinear forms

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{p} T_{ijk} x_i y_j z_k.$$

The rank of a matrix A is the smallest number r such that A can be written as the sum of r outer products xy'. We define the *tensor rank* of a tensor analogously. A *rank one tensor* is an outer product of

the form $\sum_{i,j,k} x_i y_j z_k$. The *tensor rank* of a tensor T is the smallest number r such that T can be written as the sum of r rank one tensors:

$$T = \sum_{s=1}^{r} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{p} x_i^{(s)} y_j^{(s)} z_k^{(s)}.$$

We denote the rank of a tensor T by $R(T)$. In contrast to matrix rank, tensor rank is NP-hard to compute [6].

The *matrix multiplication tensor* $\langle n, m, p \rangle$ is given by

$$T = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{p} x_{ij} y_{jk} z_{ki}.$$

This is an $nm \times mp \times pn$ tensor which corresponds to the trilinear product $\mathrm{Tr}(xyz)$, where x, y, z are interpreted as $n \times m, m \times p, p \times n$ matrices, correspondingly. Strassen [10] proved that

$$\omega = \lim_{n \to \infty} R(\langle n, n, n \rangle)^{1/n}.$$

The *volume* of a matrix multiplication tensor $\langle n, m, p \rangle$ is defined to be $\mathrm{Vol}(\langle n, m, p \rangle) = nmp$. This is also the number of non-zero entries in the corresponding three-dimensional array.

If A_i is a sequence of matrices converging to a matrix A, then $R(A_i) \to R(A)$. The same doesn't hold for tensors: all we are guaranteed is that $\lim_i R(A_i) \leq R(A)$. For example, the following tensor (appearing in [1]) has rank 3 if $\epsilon = 0$ and rank 2 if $\epsilon \neq 0$:

$$\begin{pmatrix} 1 & 0 \\ \epsilon & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

The *border rank* of a tensor T, denoted $\underline{R}(T)$, is the smallest rank of a sequence of tensors converging to T. Equivalently, the border rank of T is the smallest rank over $\mathbb{R}[\epsilon]$ of any tensor of the form $\epsilon^k T + \sum_{\ell=k+1}^{r} \epsilon^\ell T_\ell$ (the equivalence is not immediate but follows from a result of Strassen [12], see [2, §20.6]). We denote any tensor of the latter form by $\epsilon^k T + O(\epsilon^{k+1})$.

Two tensors are *equivalent* if they differ by a permutation of the rows, columns and "stacks" (independently). We denote equivalence by $\approx$. Oftentimes we think of tensors as being defined only up to equivalence.

For matrices A_1, A_2 of dimensions $n_1 \times m_1, n_2 \times m_2$, their direct sum $A_1 \oplus A_2$ is the $(n_1 + n_2) \times (m_1 + m_2)$ block-diagonal matrix having as blocks A_1, A_2. Similarly we can define the *direct sum* of two tensors T_1, T_2.

The Kronecker or tensor product of matrices is a less familiar operation: the tensor product $A_1 \otimes A_2$ is an $n_1 n_2 \times m_1 m_2$ matrix whose entries are $(A_1 \otimes A_2)_{i_1 i_2, j_1 j_2} = (A_1)_{i_1, j_1} (A_2)_{i_2, j_2}$. The *tensor product* of two tensors is defined analogously. It then follows immediately that $\langle n_1, m_1, p_1 \rangle \otimes \langle n_2, m_2, p_2 \rangle \approx \langle n_1 n_2, m_1 m_2, p_1 p_2 \rangle$. The nth *tensor power* of a tensor T is denoted by $T^{\otimes n}$. Both rank and border rank are submultiplicative: $R(T_1 \otimes T_2) \leq R(T_1) R(T_2)$ and $\underline{R}(T_1 \otimes T_2) \leq \underline{R}(T_1) \underline{R}(T_2)$.

The *support* of a tensor is the set of variables (corresponding to rows, columns and "stacks") appearing in the corresponding trilinear product.

Matrices can be transposed. The corresponding operation for tensors is *rotation*. For an $n \times m \times p$ tensor $T = \sum_{ijk} T_{ijk} x_i y_j z_k$, its rotation is the $m \times p \times n$ tensor $T' = \sum_{jki} T_{ijk} y_j z_k x_i$. Repeating the operation again, we obtain a $p \times n \times m$ tensor T''. All rotations of a tensor have the same rank and the same border rank.

Schönhage [8] proved the following fundamental theorem, which is the main vehicle used for proving upper bounds on ω.

Theorem 2.1 (Asymptotic sum inequality)**.** *For every set n_i, m_i, p_i $(1 \leq i \leq K)$ of positive integers,*

$$\sum_{i=1}^{K} (n_i m_i p_i)^{\omega/3} \leq \underline{R} \left(\bigoplus_{i=1}^{K} \langle n_i, m_i, p_i \rangle \right).$$

2.2 Laser method

The material below is adapted from [2, Chapter 15] and the papers [5, 7].

The asymptotic sum inequality deduces a bound on ω from a bound on the border rank of a direct sum of matrix multiplication tensors. The laser method, due to Strassen [12], is a technique that exploits a bound on the border rank of a sum of *non-disjoint* matrix multiplication tensors:

$$\underline{R}\left(\sum_{s=1}^{K} T_s\right) \leq r.$$

The tensors $T_1, \ldots, T_K$, known as *constituent tensors*, are trilinear forms in variables x_i, y_j, z_k, which we call x-variables, y-variables and z-variables:

$$T_s = \sum_{i,j,k} T_{s,i,j,k} x_i y_j z_k.$$

The x-variables, y-variables and z-variables are each partitioned into *blocks*. The support of each tensor is constrained to be a union of blocks. The sum $\sum_s T_s$ together with the partitions of the variables constitute a *partitioned tensor*. Two constituent tensors of a partitioned tensor are *disjoint* if their support is disjoint.

If $T = \sum_s T_s$ and $V = \sum_u V_u$ are two partitioned tensors, then $T \otimes V = \sum_{s,u} T_s \otimes V_u$ is the partitioned tensor obtained by taking the "product" partitions, that is, if x_i belongs to block X_I in T and x_j belongs to block X_J in V then x_{ij} belongs to block X_{IJ} in $T \otimes V$.

If $T = \sum_s T_s$ then $T' = \sum_s T'_s$ is the partitioned tensor obtained by rotating the constitutent tensors and retaining the partitions.

The concept of *value* plays a central role.

Definition 2.1. Let T be a partitioned tensor whose constituent tensors are matrix multiplication tensors. A *zeroing degeneration* of T is any tensor obtained from T by zeroing blocks of variables (i.e., deleting rows, columns and "stacks").

Let T be a partitioned tensor satisfying $T' = T$ (we say that T is *symmetric*). For $\rho \in [2, 3]$ and $N \geq 1$, we define $V_{\rho,N}(T)$ to be the maximum of $\sum_{i=1}^{L} (n_i m_i p_i)^{\rho/3}$ over all zeroing degenerations of $T^{\otimes N}$ with disjoint constituent tensors equivalent to $\langle n_1, m_1, p_1 \rangle, \ldots, \langle n_L, m_L, p_L \rangle$. The *value* of T is the function

$$V_\rho(T) = \lim_{n \to \infty} V_{\rho,N}(T)^{1/N}.$$

The existence of the limit was proved by Stothers [9, 5].

For an arbitrary partitioned tensor T, we define

$$V_\rho(T) = V_\rho(T \otimes T' \otimes T'')^{1/3}.$$

Our definition is different from the one appearing in [9, 13, 5, 7]. They define $V_{\rho,N}(T)$ to be the maximum of $\sum_{i=1}^{L} (n_i m_i p_i)^{\rho/3}$ over all zeroing degenerations of $T^{\otimes N}$ equivalent to $\bigoplus_{i=1}^{L} \langle n_i, m_i, p_i \rangle$. However, as we detail in Section 3.2, their actual usage of the definition is more constrained (though more general than Definition 2.1). The advantage of our definition is that in some cases it is possible to *compute* the value.

For completeness, we include the proof that $V_\rho(T)$ is well-defined.

Lemma 2.2. *Let T be a symmetric partitioned tensor. The limit $\lim_{n\to\infty} V_{\rho,N}(T)^{1/N}$ exists.*

Proof. We start by showing that

$$V_{\rho,N_1+N_2}(T) \geq V_{\rho,N_1}(T) V_{\rho,N_2}(T).$$

Indeed, let S_1, S_2 be the zeroing degenerations of $T^{\otimes N_1}, T^{\otimes N_2}$ yielding $V_{\rho,N_1}(T), V_{\rho,N_2}(T)$, say $S_1 = \sum_i S_{1,i}$ with $V_{\rho,N_1}(T) = \sum_i \mathrm{Vol}(S_{1,i})^{\rho/3}$ and $S_2 = \sum_j S_{2,j}$ with $V_{\rho,N_2}(T) = \sum_j \mathrm{Vol}(S_{2,j})^{\rho/3}$. It is not hard to check

that $S_1 \otimes S_2$ is a zeroing degeneration of $T^{\otimes(N_1+N_2)}$ with disjoint constituent tensors $S_1 \otimes S_2 = \sum_{i,j} S_{1,i} \otimes S_{2,j}$, and so

$$V_{\rho,N_1+N_2}(T) \geq \sum_{i,j} \mathrm{Vol}(S_{1,i} \otimes S_{2,j})^{\rho/3} = \sum_{i,j} \mathrm{Vol}(S_{1,i})^{\rho/3} \mathrm{Vol}(S_{2,j})^{\rho/3} = V_{\rho,N_1}(T) V_{\rho,N_2}(T).$$

Hence the sequence $N \mapsto \log_2 V_{\rho,N}(T)$ is superadditive, and so Fekete's lemma shows that $(\log_2 V_{\rho,N}(T))/N = \log_2 V_{\rho,N}^{1/N}$ tends to a limit. □

Any matrix multiplication tensor constitutes a paritioned tensor with a unique constituent tensor. The following formula follows directly from the definition.

Lemma 2.3. *The value of a matrix multiplication tensor $\langle n, m, p \rangle$ is $V_\rho(\langle n, m, p \rangle) = (nmp)^{\rho/3}$.*

The laser method as originally conceived by Strassen used a more powerful concept of degeneration related to the border rank; see [2, 15.6] for more details.

Stothers [9, 5] extended the asymptotic sum inequality to the setting of partitioned tensors.

Theorem 2.4. *For partitioned tensors $T_1, \ldots, T_K$ whose constituent tensors are matrix multiplication tensors,*

$$\sum_{i=1}^{K} V_\omega(T_i) \leq \underline{R}\left(\bigoplus_{i=1}^{K} T_i\right).$$

Proof. Consider first the case of a single partitioned tensor T which is symmetric. The asymptotic sum inequality shows that $V_{\omega,N}(T)^{1/N} \leq \underline{R}(T^{\otimes N})^{1/N} \leq \underline{R}(T)$. Taking the limit $N \to \infty$ completes the proof. Consider next the case of an asymmetric single partitioned tensor T. The tensor $T \otimes T' \otimes T''$ is symmetric and so $V_\omega(T) = V_\omega(T \otimes T' \otimes T'')^{1/3} \leq \underline{R}(T \otimes T' \otimes T'')^{1/3} \leq \underline{R}(T)$, since $\underline{R}(T) = \underline{R}(T') = \underline{R}(T'')$.

In order to complete the proof, we show that $V_\omega(T_1 \oplus T_2) \geq V_\omega(T_1) + V_\omega(T_2)$. Suppose first that T_1, T_2 are symmetric. The tensor $(T_1 \oplus T_2)^{\otimes N}$ decomposes as

$$(T_1 \oplus T_2)^{\otimes N} = \bigoplus_{N_1+N_2=N} \binom{N}{N_1} \odot T_1^{\otimes N_1} T_2^{\otimes N_2},$$

where $M \odot T$ denotes the direct sum of M tensors equivalent to T. In particular, as in the proof of Lemma 2.2,

$$V_{\rho,N}(T_1 \oplus T_2) \geq \sum_{N_1+N_2=N} \binom{N}{N_1} V_{\rho,N_1}(T_1) V_{\rho,N_2}(T_2).$$

Let $\alpha_1 = V_\rho(T_1)/(V_\rho(T_1) + V_\rho(T_2))$ and $\alpha_2 = V_\rho(T_2)/(V_\rho(T_1) + V_\rho(T_2))$. Considering $N_1 \approx \alpha_1 N$ and $N_2 \approx \alpha_2 N$, we get

$$V_{\rho,N}(T_1 \oplus T_2) \gtrapprox 2^{H(\alpha_1,\alpha_2)N} V_\rho(T_1)^{\alpha_1 N} V_\rho(T_2)^{\alpha_2 N} = (V_\rho(T_1) + V_\rho(T_2))^N,$$

where the approximation is true up to polynomial factors and in the limit $N \to \infty$. This shows that $V_\rho(T_1 \oplus T_2) \geq V_\rho(T_1) + V_\rho(T_2)$. □

2.3 Coppersmith–Winograd identity

Coppersmith and Winograd [4] exhibit the following identity, for any $q \geq 1$:

$$\epsilon^3 \left[\sum_{i=1}^{q} \left(x_0^{[0]} y_i^{[1]} z_i^{[1]} + x_i^{[1]} y_0^{[0]} z_i^{[1]} + x_i^{[1]} y_i^{[1]} z_0^{[0]} \right) + x_0^{[0]} y_0^{[0]} z_{q+1}^{[2]} + x_0^{[0]} y_{q+1}^{[2]} z_0^{[0]} + x_{q+1}^{[2]} y_0^{[0]} z_0^{[0]} \right] + O(\epsilon^4) =$$

$$\epsilon \sum_{i=1}^{q} (x_0^{[0]} + \epsilon x_i^{[1]})(y_0^{[0]} + \epsilon y_i^{[1]})(z_0^{[0]} + \epsilon z_i^{[1]}) -$$

$$\left(x_0^{[0]} + \epsilon^2 \sum_{i=1}^{q} x_i^{[1]} \right) \left(y_0^{[0]} + \epsilon^2 \sum_{i=1}^{q} y_i^{[1]} \right) \left(z_0^{[0]} + \epsilon^2 \sum_{i=1}^{q} z_i^{[1]} \right) +$$

$$(1 - q\epsilon)(x_0^{[0]} + \epsilon^3 x_{q+1}^{[2]})(y_0^{[0]} + \epsilon^3 y_{q+1}^{[2]})(z_0^{[0]} + \epsilon^3 z_{q+1}^{[2]}).$$

Here the superscripts denote the partitions:

$$x_0^{[0]}; x_1^{[1]}, \ldots, x_q^{[1]}; x_{q+1}^{[2]},$$

and similarly for the y-variables and z-variables. We denote this partition $X^{[0]}, X^{[1]}, X^{[2]}$. The identity shows that

$$\underline{R}(\langle 1,1,q \rangle^{0,1,1} + \langle q,1,1 \rangle^{1,0,1} + \langle 1,q,1 \rangle^{1,1,0} + \langle 1,1,1 \rangle^{0,0,2} + \langle 1,1,1 \rangle^{0,2,0} + \langle 1,1,1 \rangle^{2,0,0}) \leq q + 2.$$

Here $\langle n,m,p \rangle^{I,J,K}$ denotes a matrix multiplication tensor equivalent to $\langle n,m,p \rangle$ whose support is $X^{[I]}, Y^{[J]}, Z^{[K]}$. We denote the corresponding partitioned tensor T_{CW} (depending on q).

Coppersmith and Winograd calculated the value of T_{CW}.

Theorem 2.5. *For $q \geq 1$,*

$$\log_2 V_\rho(T_{CW}) = \max_{0 \leq \alpha \leq 1} H(\tfrac{2-\alpha}{3}, \tfrac{2\alpha}{3}, \tfrac{1-\alpha}{3}) + \tfrac{1}{3}\rho\alpha \log_2 q.$$

Proof. The upper bound appears implicitly in Cohn, Kleinberg, Szegedy and Umans [3, Lemma 3.2], but since they discuss a slightly different case, we proceed to spell it out.

The tensor T_{CW} is symmetric, and so

$$\log_2 V_\rho(T_{CW}) = \lim_{n \to \infty} \frac{\log_2 V_{\rho,N}(T_{CW})}{N}.$$

Suppose that S is a zeroing degeneration of $T^{\otimes N}$ which is the sum of disjoint constituent tensors $\langle n_i, m_i, p_i \rangle$. Our goal is to bound $\sum_i (n_i m_i p_i)^{\rho/3}$.

The tensor $T_{CW}^{\otimes N}$ has 6^N constituent tensors, each resulting from the tensor product of N original constituent tensors. The *type* of a constitutent tensor of $T_{CW}^{\otimes N}$ is $(\alpha_x, \alpha_y, \alpha_z, \beta_x, \beta_y, \beta_z)$ if the tensor resulted from multiplying $\alpha_x N, \alpha_y N, \alpha_z N, \beta_x N, \beta_y N, \beta_z N$ copies each of $\langle 1,1,q \rangle^{0,1,1}, \langle q,1,1 \rangle^{1,0,1}, \langle 1,q,1 \rangle^{1,1,0}$, $\langle 1,1,1 \rangle^{2,0,0}, \langle 1,1,1 \rangle^{0,2,0}, \langle 1,1,1 \rangle^{0,0,2}$, respectively. Note that there are $O(N^5)$ many types.

Each constituent tensor in $T_{CW}^{\otimes N}$ has an associated block of x-variables, which corresponds to a vector in $\{0,1,2\}^N$. The *x-type* of this tensor is the fraction of coordinates of each type, which we denote by (c_0, c_1, c_2). The y-type and z-type are defined analogously.

Let $\tau = (\alpha_x, \alpha_y, \alpha_z, \beta_x, \beta_y, \beta_z)$ be any type. Each constitutent tensor of type τ has volume $q^{(\alpha_x+\alpha_y+\alpha_z)N}$, x-type $(\alpha_x + \beta_y + \beta_z, \alpha_y + \alpha_z, \beta_x)$, y-type $(\alpha_y + \beta_x + \beta_z, \alpha_x + \alpha_z, \beta_y)$ and z-type $(\alpha_z + \beta_x + \beta_y, \alpha_x + \alpha_y, \beta_z)$. Since the constituent tensors of type τ are disjoint, in particular they have distinct x-types. Since they all have the same volume, we conclude that

$$\begin{aligned} \sum_{i \in \tau} (n_i m_i p_i)^{\rho/3} &\leq q^{(\alpha_x+\alpha_y+\alpha_z)(\rho/3)N} \binom{N}{(\alpha_x + \beta_y + \beta_z)N, (\alpha_y + \alpha_z)N, \beta_x N} \\ &\leq q^{(\alpha_x+\alpha_y+\alpha_z)N} \exp_2 H(\alpha_x + \beta_y + \beta_z, \alpha_y + \alpha_z, \beta_x)N. \end{aligned}$$

We can similarly get a bound from y-types and z-types. Let $\alpha = \alpha_x + \alpha_y + \alpha_z$, and notice that $\beta_x + \beta_y + \beta_z = 1 - \alpha$. Taking the geometric mean of these bounds, we obtain

$$\begin{aligned}
&\sum_{i \in \tau} (n_i m_i p_i)^{\rho/3} \\
\leq & q^{\alpha(\rho/3)N} \exp_2 \frac{H(\alpha_x + \beta_y + \beta_z, \alpha_y + \alpha_z, \beta_x) + H(\alpha_y + \beta_x + \beta_z, \alpha_x + \alpha_z, \beta_y) + H(\alpha_z + \beta_x + \beta_y, \alpha_x + \alpha_y, \beta_z)}{3} N \\
\leq & q^{\alpha(\rho/3)N} \exp_2 H\left(\frac{2-\alpha}{3}, \frac{2\alpha}{3}, \frac{1-\alpha}{3}\right) N,
\end{aligned}$$

using concavity of the entropy function. Summing over all $O(N^5)$ different types, we obtain

$$V_{\rho,N}(T_{CW}) \leq O(N^5) \max_{\alpha \in [0,1]} q^{\alpha(\rho/3)N} \exp_2 H\left(\frac{2-\alpha}{3}, \frac{2\alpha}{3}, \frac{1-\alpha}{3}\right) N,$$

and so

$$\log_2 V_\rho(T_{CW}) \leq \max_{\alpha \in [0,1]} H(\tfrac{2-\alpha}{3}, \tfrac{2\alpha}{3}, \tfrac{1-\alpha}{3}) + \tfrac{1}{3}\rho\alpha \log_2 q.$$

This completes the proof of the upper bound.

The lower bound on $\log_2 V_\rho(T_{CW})$ follows from the work of Coppersmith and Winograd [4], and appears more formally in [5, 7]; see in particular [7, Theorem 4.1]. The idea is as follows. Given α and N, let τ be a realizable type close to $(\alpha/3, \alpha/3, \alpha/3, (1-\alpha)/3, (1-\alpha)/3, (1-\alpha)/3)$. Every constituent tensor of this type has x-type, y-type and z-type $((2-\alpha)/3, 2\alpha/3, (1-\alpha)/3)$. Coppersmith and Winograd start by zeroing out all x, y, z-variables not of this x, y, z-type (respectively). Then they zero out more variables to obtain $\exp_2(H(\frac{2-\alpha}{3}, \frac{2\alpha}{3}, \frac{1-\alpha}{3}) - o(1))N$ *disjoint* constituent tensors (this is the difficult part of the construction). This shows that $V_{\rho,N}(T_{CW}) \geq q^{\alpha(\rho/3)N} \exp_2(H(\frac{2-\alpha}{3}, \frac{2\alpha}{3}, \frac{1-\alpha}{3}) - o(1))N$, and the lower bound follows by taking the limit $N \to \infty$. □

Routine calculation reveals that the optimal α given q, ρ is

$$\alpha = \frac{-3 + \sqrt{32q^{-\rho} + 1}}{8q^{-\rho} - 2},$$

from which we can get an explicit expression for $V_\rho(T_{CW})$. Following Coppersmith and Winograd, we plug $q = 6$ and use Theorem 2.4 (with the single partitioned tensor T_{CW}) to obtain $\omega < 2.38719$.

In their paper, Coppersmith and Winograd introduced another technique which enabled them to obtain an improved bound $\omega < 2.375477$. The technique, which we call *merging*, was used again by Vassilevska-Williams [13], Stothers [9] and Le Gall [7], culminating in the bound $\omega < 2.3728639$ obtained by Le Gall. Our goal in the rest of the paper is to describe this technique and its limitations.

Remark Le Gall [7, Theorem 4.1], following Stothers and Vassilevska-Williams, describes a more general situation and obtains a general lower bound on the value of general partitioned tensors. The argument of Theorem 2.5 yields an almost matching upper bound, differing only in the ommission of the term $\Gamma_S(P)$. This term is not present in the calculations of Coppersmith and Winograd, but appears in Stothers' and subsequent work of Vassilevska-Williams and Le Gall. As a consequence, we do not know whether the constructions described in Stothers' and subsequent work are tight with respect to the technique they use. It is an open question to close this gap. As an aside, we note that even if this question is settled, there remains a difficult numerical optimization problem which is solved only approximately in these papers; in particular, the parameters used for the construction are only approximately optimal.

3 Merging

3.1 Definition

We start with a formal definition of merging.

Definition 3.1. Let T be a symmetric partitioned tensor. For $\rho \in [2,3]$ and $N \geq 1$, we define $V^*_{\rho,N}(T)$ to be the maximum of $\sum_{i=1}^{L}(n_i m_i p_i)^{\omega/3}$ over all zeroing degenerations of $T^{\otimes N}$ with sets $T_1, \ldots, T_L$ of constitutent tensors on disjoint variables such that $\sum T_i \coloneqq \sum_{U \in T_i} U$ is equivalent to $\langle n_i, m_i, p_i \rangle$. The *extended value* of T is the function

$$V^*_\rho(T) = \lim_{n\to\infty} V^*_{\rho,N}(T)^{1/N}.$$

(We prove the existence of the limit below.)

For an arbitrary partitioned tensor T, we define

$$V^*_\rho(T) = V^*_\rho(T \otimes T' \otimes T'')^{1/3}.$$

The constituent tensors in each T_i are merged together to one big matrix multiplication tensor, whence the name *merging*. Before giving an example, let us prove that $V^*_\rho(T)$ is well-defined.

Lemma 3.1. *Let T be a symmetric partitioned tensor and $\rho \in [2,3]$. The limit $\lim_{n\to\infty} V^*_{\rho,N}(T)^{1/n}$ exists.*

Proof. The proof is very similar to the proof of Lemma 2.2. It is enough to prove the inequality

$$V^*_{\rho,N_1+N_2}(T) \geq V^*_{\rho,N_1}(T) V^*_{\rho,N_2}(T),$$

and then the existence of the limit follows from Fekete's lemma. Let S_1, S_2 be zeroing degenerations of $T^{\otimes N_1}, T^{\otimes N_2}$ witnessing $V^*_{\rho,N_1}(T), V^*_{\rho,N_2}(T)$, that is, $S_1 = \sum_i \sum S_{1,i}$ with $\sum_i \operatorname{Vol}(\sum S_{1,i})^{\rho/3} = V^*_{\rho,N_1}(T)$, and $S_2 = \sum_j \sum S_{2,j}$ with $\sum_j \operatorname{Vol}(\sum S_{2,j})^{\rho/3} = V^*_{\rho,N_2}(T)$. The tensor $S_1 \otimes S_2$ is a zeroing degeneration of $T^{\otimes(N_1+N_2)}$ and $S_1 \otimes S_2 = \sum_{i,j} \sum S_{1,i} \otimes S_{2,j}$. Since $\operatorname{Vol}(S_{1,i} \otimes S_{2,j}) = \operatorname{Vol}(S_{1,i}) \operatorname{Vol}(S_{2,j})$, we obtain $V^*_{\rho,N_1+N_2}(T) \geq \sum_{i,j} \operatorname{Vol}(S_{1,i} \otimes S_{2,j})^{\rho/3} = V^*_{\rho,N_1} V^*_{\rho,N_2}$. □

The proof of Theorem 2.4 extends *mutatis mutandis*.

Theorem 3.2. *For partitioned tensors $T_1, \ldots, T_K$ whose constituent tensors are matrix multiplication tensors,*

$$\sum_{i=1}^{K} V^*_\omega(T_i) \leq \underline{R}\left(\bigoplus_{i=1}^{K} T_i\right).$$

3.2 Relation to known constructions

As an illustration of the concept of merging, we present the bound $\omega < 2.375477$ obtained by Coppersmith and Winograd [4]. They consider the tensor square $T^{\otimes 2}_{CW}$, repartitioning it according to the sum of the indices of the original blocks:

$$X'_0 = X_{00}, X'_1 = X_{01} \cup X_{10}, X'_2 = X_{20} \cup X_{11} \cup X_{02}, X'_3 = X_{21} \cup X_{12}, X'_4 = X_{22}.$$

The resulting partitioned tensor $\tilde{T}^{\otimes 2}_{CW}$ has 15 constituent tensors:

(a) 3 constituent tensors equivalent to $T_1 = \langle 1,1,1 \rangle^{0,0,4}$, coming from

$$\langle 1,1,1 \rangle^{0,0,2} \otimes \langle 1,1,1 \rangle^{0,0,2}.$$

(b) 6 constituent tensors equivalent to $T_2 = \langle 1,1,2q \rangle^{0,1,3}$ and its permutations $T'_2, \ldots, T^{(5)}_2$, coming from

$$\langle 1,1,q \rangle^{0,1,1} \otimes \langle 1,1,1 \rangle^{0,0,2} \oplus \langle 1,1,1 \rangle^{0,0,2} \otimes \langle 1,1,q \rangle^{0,1,1}.$$

(c) 3 constituent tensors equivalent to $T_3 = \langle 1,1,q^2+2\rangle^{0,2,2}$ and its rotations T_3', T_3'', coming from

$$\langle 1,1,1\rangle^{0,2,0} \otimes \langle 1,1,1\rangle^{0,0,2} \oplus \langle 1,1,1\rangle^{0,0,2} \otimes \langle 1,1,1\rangle^{0,2,0} \oplus \langle 1,1,q\rangle^{0,1,1} \otimes \langle 1,1,q\rangle^{0,1,1}.$$

(d) 3 constituent tensors equivalent to $T_4^{1,1,2} = S_1 + S_2 + S_3 + S_4$ and its rotations T_4', T_4'', coming from

$$\langle 1,q,1\rangle^{1,1,0} \otimes \langle 1,1,1\rangle^{0,0,2} + \langle 1,1,1\rangle^{0,0,2} \otimes \langle 1,q,1\rangle^{1,1,0} + \langle q,1,1\rangle^{1,0,1} \otimes \langle 1,1,q\rangle^{0,1,1} + \langle 1,1,q\rangle^{0,1,1} \otimes \langle q,1,1\rangle^{1,0,1},$$

where $S_1 = \langle 1,q,1\rangle^{10,10,02}$, $S_2 = \langle 1,q,1\rangle^{01,01,20}$, $S_3 = \langle q,1,q\rangle^{10,01,11}$, $S_4 = \langle q,1,q\rangle^{01,10,11}$.

Coppersmith and Winograd compute the value $V_\rho(T_4)$ directly. They proceed to compute a generalized form of the value of $\tilde{T}_{CW}^{\otimes 2}$, obtained by pretending that T_4 is a matrix multiplication tensor with $\mathrm{Vol}(T_4)^{\rho/3} = V_\rho(T_4)$, and use a corresponding generalization of Theorem 2.4 to obtain an upper bound on ω:

$$V_\omega(\tilde{T}_{CW}^{\otimes 2}) \leq \underline{R}(\tilde{T}_{CW}^{\otimes 2}) \leq \underline{R}(T_{CW})^2.$$

The proof of the generalized version relies on the property $V_{\rho,N}(T_4 \otimes T_4' \otimes T_4'') \approx V_\rho(T_4)^{3N}$, and is otherwise identical to the proof of Theorem 2.4.

We proceed to recast their construction in terms of the extended value. For given N, the first step of their construction is taking the Nth power of $\tilde{T}_{CW}^{\otimes 2}$ and zeroing variables to obtain a certain quantity X_N of constituent tensors equivalent to

$$T_1^{\otimes \alpha_1 N} \otimes (T_2 \otimes \cdots \otimes T_2^{(5)})^{\otimes (\alpha_2/6)N} \otimes (T_3 \otimes T_3' \otimes T_3'')^{\otimes (\alpha_3/3)N} \otimes (T_4 \otimes T_4' \otimes T_4'')^{\otimes (\alpha_4/3)N},$$

for some parameters $\alpha_1, \alpha_2, \alpha_3, \alpha_4$. Attention is then focused on the factor $(T_4 \otimes T_4' \otimes T_4'')^{\otimes (\alpha_4/3)N}$. More variables are zeroed to obtain a certain quantity $Y_{(\alpha_4/3)N}$ of constituent tensors equivalent to

$$(S_1 \otimes S_1' \otimes S_1'')^{\otimes (\beta_1/3)(\alpha_4/3)N} \otimes (S_2 \otimes S_2' \otimes S_2'')^{\otimes (\beta_2/3)(\alpha_4/3)N} \otimes (S_3 \otimes S_3' \otimes S_3'')^{\otimes (\beta_3/3)(\alpha_4/3)N} \otimes (S_4 \otimes S_4' \otimes S_4'')^{\otimes (\beta_4/3)(\alpha_4/3)N}.$$

Altogether, we obtain $X_N Y_{(\alpha_4/3)N}$ tensors equivalent to

$$T_1^{\otimes \alpha_1 N} \otimes (T_2 \otimes \cdots \otimes T_2^{(5)})^{\otimes (\alpha_2/6)N} \otimes (T_3 \otimes T_3' \otimes T_3'')^{\otimes (\alpha_3/3)N} \otimes$$
$$(S_1 \otimes S_1' \otimes S_1'')^{\otimes (\beta_1/3)(\alpha_4/3)N} \otimes (S_2 \otimes S_2' \otimes S_2'')^{\otimes (\beta_2/3)(\alpha_4/3)N} \otimes (S_3 \otimes S_3' \otimes S_3'')^{\otimes (\beta_3/3)(\alpha_4/3)N} \otimes (S_4 \otimes S_4' \otimes S_4'')^{\otimes (\beta_4/3)(\alpha_4/3)N}.$$

The tensors T_1, S_1, S_2, S_3, S_4 are constituent tensors of $T_{CW}^{\otimes 2}$, and the tensors T_2, T_3 result from *merging* constituent tensors of $T_{CW}^{\otimes 2}$. Hence this construction gives a lower bound on $V_{\rho,2N}^*(T_{CW})$ matching the lower bound obtained on the generalized value $V_{\rho,N}(\tilde{T}_{CW}^{\otimes 2})$. Taking the limit $N \to \infty$, we obtain a lower bound on $V_\rho^*(T_{CW})$ mathcing the lower bound on the generalized value $\sqrt{V_\rho(\tilde{T}_{CW}^{\otimes 2})}$. Applying Theorem 3.2, we obtain the same upper bound on ω.

Stothers [9, 5], Vassilevska-Williams [13] and Le Gall [7] generalized the ideas of Coppersmith and Winograd even further, by applying their construction recursively. Stothers considered the tensor square $\tilde{T}_{CW}^{\otimes 4}$ of $\tilde{T}_{CW}^{\otimes 2}$, again repartitioned along similar lines. We can again make a list of the constituent tensors. Some of these are matrix multiplication tensors resulting from merging constitutent tensors of $(\tilde{T}_{CW}^{\otimes 2})^{\otimes 2}$, which ultimately result from merging constitutent tensors of $T_{CW}^{\otimes 4}$. Others are complicated tensors like the tensor T_4 considered above. Stothers computes the generalized value along the lines considered above, and then applies the generalized Theorem 2.4. His construction therefore also gives a lower bound on the extended value $V_\rho^*(T_{CW})$. The constructions of Vassilevska-Williams and Le Gall proceed in the same way to analyze higher powers: Vassilevska-Williams analyzes $\tilde{T}_{CW}^{\otimes 8}$, and Le Gall analyzes $\tilde{T}_{CW}^{\otimes 16}$ and $\tilde{T}_{CW}^{\otimes 32}$. All of these results correspond to lwoer bounds on the extended value $V_\rho^*(T_{CW})$.

4 Lower bound

In this section we prove an upper bound on $V_\rho^*(T_{CW})$. This correspond to a lower bound on the value of ω which can be obtained using the technique of merging and the particular tensor T_{CW}.

We start by characterizing the possible mergings of tensors.

Lemma 4.1. *Let $q \geq 2$, and suppose that S is a subset of the constitutent tensors $T_{CW}^{\otimes N}$ such that $\sum S \approx \langle n, m, p\rangle$. Then there is a partition $[n] = X \cup Y \cup Z$ such that the x, y, z-indices A, B, C (known collectively as an* index triple*) of any tensor in S satisfy $A_t = 0$ for $t \in X$, $B_t = 0$ for $t \in Y$, and $C_t = 0$ for $t \in Z$. These are called x-*constant*, y-*constant* and z-*constant*, respectively.*

Proof. Recall that

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

Since $\sum S \approx \langle n, m, p\rangle$, there is a way to assign to each x-variable appearing in S some (unique) value x_{ij}, to each y-variable some value y_{jk}, and to each z-variable some value z_{ki}, so that after all the assignments we get exactly $\langle n, m, p\rangle$. Fix some such assignment.

Call a z-variable z_{ki} *t-good* if it appears in some tensor corresponding to some index triple $(A, B, C) \in S$ such that $(A_t, B_t, C_t) = (1, 1, 0)$. The tensor in S corresponding to the index triple (A_t, B_t, C_t) can be factored as $T = T_1 \otimes \langle 1, q, 1\rangle^{1,1,0} \otimes T_2$, where the displayed factor is the tth factor. The tensors T_1, T_2 involve many variables. Choose single x, y, z-variables from each. The corresponding subtensor of T (obtained by focusing on specific rows, columns and "stacks") has the form $\sum_{r=1}^{q} X_r Y_r Z$. For some k, i, $Z = z_{ki}$. For some functions $\alpha, \beta\colon [q] \to [m]$, $X_r = x_{i\alpha(r)}$ and $Y_r = y_{\beta(r)k}$. In particular, the denotations of all variables X_r are in the same "row" (first index), and the denotations of all variables Y_r are in the same "column" (second index).

Consider any z-variable z_{kI} for $I \in [n]$. Some tensor in S must include the product $x_{I\alpha(1)}y_{\alpha(1)k}z_{kI}$, say the one with index triple (A', B, C') (note that the y-index must indeed be B). Since $B_t = 1$, $(A'_t, B_t, C'_t) \in \{(1, 1, 0), (0, 1, 1)\}$. We want to show that $(A'_t, B_t, C'_t) = (1, 1, 0)$. Suppose to the contrary that $(A'_t, B_t, C'_t) = (0, 1, 1)$. As before, some subtensor of the tensor corresponding to (A', B, C') has the form $\sum_{r=1}^{q} X' Y_r Z'_r$. The same reasoning as before shows that the denotations of all variables Y_r are in the same row, contradicting our earlier observation that they must be in the same column (here it is essential that $q \geq 2$). We conclude that $C'_t = 0$ and so z_{kI} is also t-good.

Similar reasoning applies for any z-variable z_{Ki} for $K \in [p]$. Some tensor in S must include the product $x_{i\alpha(1)}y_{\alpha(1)K}z_{Ki}$, say the one with index triple (A, B', C'). Since $A_t = 1$, $(A_t, B'_t, C'_t) \in \{(1, 1, 0), (1, 0, 1)\}$. We want to rule out the latter case. If $(A_t, B'_t, C'_t) = (1, 0, 1)$ then some subtensor has the form $\sum_{r=1}^{q} X_r Y' Z'_r$. It follows that the denotations of all variables X_r are in the same column, whereas earlier we observed that they must be in the same row. Since $q \geq 2$, we obtain a contradiction, and again conclude that $C'_t = 0$ and so z_{Ki} is t-good.

We have shown that if z_{ki} is t-good then for all $I \in [n]$ and $K \in [p]$, the z-variables z_{kI}, z_{Ki} are t-good. It follows by transitivity that all variables z_{KI} are t-good, and so t is z-constant. This shows that as long as $(A_t, B_t, C_t) \in \{(1, 1, 0), (1, 0, 1), (0, 1, 1)\}$ for some index triple $(A, B, C) \in S$, the coordinate t is either z-constant, y-constant or x-constant (respectively).

It remains to consider coordinates t such that for all index triples $(A, B, C) \in S$, it is the case that $(A_t, B_t, C_t) \in \{(2, 0, 0), (0, 2, 0), (0, 0, 2)\}$. If only two of the options occur then the coordinate is z-constant (if the third doesn't appear), y-constant (if the second doesn't appear) or x-constant (if the first doesn't appear), so it remains to rule out the case in which all of these options occur.

Say that an x-variable x_{ij} has t-type $\tau \in \{0, 2\}$ if the x-index A in which its denotation appears satisfies $A_t = \tau$, and define the t-type of y-variables and z-variables analogously. By assumption, there are some variables x_{ij}, y_{pq}, z_{rs} of t-type 2. The product $x_{ip}y_{pq}z_{qi}$ corresponds to some index triple $(A, B, C) \in S$ satisfying $B_t = 2$. We conclude that $(A_t, B_t, C_t) = (0, 2, 0)$ and so x_{ip} has t-type 0. Similarly, the product $x_{sp}y_{pr}z_{rs}$ shows that y_{pr} has t-type 0, and the product $x_{ij}y_{jr}z_{ri}$ shows that z_{ri} has t-type 0. But then the

product $x_{ip}y_{pr}z_{ri}$ corresponds to some index triple $(A,B,C) \in S$ satisfying $(A_t,B_t,C_t) = (0,0,0)$, which is impossible. This contradiction completes the proof. □

Given this characterization, we can prove our main theorem.

Theorem 4.2. *For every $q \geq 2$,*

$$V^*_\rho(T_{CW}) \leq \max_{\alpha\in[0,1]} q^{\alpha\rho/3} \exp_2 H(\tfrac{2-\alpha}{3}, \tfrac{2\alpha}{3}, \tfrac{1-\alpha}{3}) \exp_2[H(\tfrac{1-\alpha}{2}, \alpha, \tfrac{1-\alpha}{2})\tfrac{\rho-2}{3}].$$

Proof. Given an integer N, we will bound $V^*_{\rho,N}(T_{CW})$. Let T be a zeroing degeneration $T_{CW}^{\otimes n}$ and $T = \bigcup_i S_i$ be a decomposition such that $S_i \approx \langle n_i, m_i, p_i\rangle$ and $V^*_{\rho,N}(T_{CW}) = \sum_i (n_i m_i p_i)^{\rho/3}$. We call each S_i a *line*. Lemma 4.1 shows that for each line S_i, each $t \in [N]$ is either x-constant, y-constant or z-constant. If there are $\gamma_x N, \gamma_y N, \gamma_z N$ of each then we say that S_i has line type $\tau_\ell = (\gamma_x, \gamma_y, \gamma_z)$. There are $O(N^2)$ different line types.

Each tensor $T \in S_i$ has an associated index triple (A,B,C), as in Lemma 4.1. We define the type of T, the x-type of A, the y-type of B and the z-type of C as in Theorem 2.5. There are $O(N^5)$ possible types. We let $\mathrm{Vol}_\tau(S_i)$ be the sum of the volumes of all $T \in S_i$ of type τ. Since the volume is the number of basic products xyz, it follows that $nmp = \mathrm{Vol}(\sum S_i) = \sum_\tau \mathrm{Vol}_\tau(S_i)$.

Consider a specific line type $\tau_\ell = (\gamma_x, \gamma_y, \gamma_z)$ and a specific type $\tau = (\alpha_x, \alpha_y, \alpha_z, \beta_x, \beta_y, \beta_z)$. We will upper bound

$$U_{\rho,N}(\tau_\ell, \tau) = \sum_{i\colon S_i \text{ has type } \tau_\ell} \mathrm{Vol}_\tau(S_i)^{\rho/3}.$$

This implies an upper bound on $V^*_{\rho,N}(T_{CW})$ as follows. First,

$$\begin{aligned}\sum_{i\colon S_i \text{ has type } \tau_\ell} \mathrm{Vol}(S_i)^{\rho/3} &= \sum_{i\colon S_i \text{ has type } \tau_\ell} \left(\sum_\tau \mathrm{Vol}_\tau(S_i)\right)^{\rho/3} \\ &\leq \sum_{i\colon S_i \text{ has type } \tau_\ell} \left(O(N^5) \max_\tau \mathrm{Vol}_\tau(S_i)\right)^{\rho/3} \leq O(N^5) \max_\tau U_{\rho,N}(\tau_\ell, \tau).\end{aligned}$$

Summing over all τ_ℓ,

$$V^*_{\rho,N}(T_{CW}) \leq O(N^7) \max_{\tau_\ell,\tau} U_{\rho,N}(\tau_\ell, \tau).$$

When taking the Nth root and letting $N \to \infty$, the factor $O(N^7)$ disappears. Therefore

$$V^*(T_{CW}) \leq \max_{\tau_\ell,\tau} \lim_{N\to\infty} U_{\rho,N}(\tau_\ell, \tau)^{1/N}. \tag{1}$$

Let $\alpha = \alpha_x + \alpha_y + \alpha_z$ and $\beta = \beta_x + \beta_y + \beta_z$, and define

$$\begin{aligned}
P_x &= \exp_2 H(\alpha_x + \beta_y + \beta_z, \alpha_y + \alpha_z, \beta_x)N, \\
P_y &= \exp_2 H(\beta_x + \alpha_y + \beta_z, \alpha_x + \alpha_z, \beta_y)N, \\
P_z &= \exp_2 H(\beta_x + \beta_y + \alpha_z, \alpha_x + \alpha_y, \beta_z)N, \\
Q_x &= \exp_2 H\big(\tfrac{\alpha_x+\beta_y+\beta_z-\gamma_x}{1-\gamma_x}, \tfrac{\alpha_y+\alpha_z}{1-\gamma_x}, \tfrac{\beta_x}{1-\gamma_x}\big)(1-\gamma_x)N, \\
Q_y &= \exp_2 H\big(\tfrac{\beta_x+\alpha_y+\beta_z-\gamma_y}{1-\gamma_y}, \tfrac{\alpha_x+\alpha_z}{1-\gamma_y}, \tfrac{\beta_y}{1-\gamma_y}\big)(1-\gamma_y)N, \\
Q_z &= \exp_2 H\big(\tfrac{\beta_x+\beta_y+\alpha_z-\gamma_z}{1-\gamma_z}, \tfrac{\alpha_x+\alpha_y}{1-\gamma_z}, \tfrac{\beta_z}{1-\gamma_z}\big)(1-\gamma_z)N.
\end{aligned}$$

Here P_x, P_y, P_z are upper bounds on the number of different x,y,z-indices, respectively. The quantities Q_x, Q_y, Q_z are upper bounds on the number of different x,y,z–indices, respectively, that can appear in any given line. The reason that Q_x bounds the number of x-indices is that a γ_x-fraction of the indices are fixed at

0, and these have to be deducted from the $\alpha_x + \beta_y + \beta_z$-fraction which is 0 among the entire N coordinates. The resulting distribution then applies only to the remaining $(1-\gamma_x)N$ coordinates.

From now on, we consider only lines of line type τ_ℓ. Let I_t, J_t, K_t be the number of x, y, z-indices, respectively, in tensors of type τ in line S_t. Note that $\sum_t I_t \leq P_x$, $\sum_t J_t \leq P_y$, $\sum_t K_t \leq P_z$. As noted above, $I_t \leq Q_x$, $J_t \leq Q_y$, $K_t \leq Q_z$. In order to upper bound $\mathrm{Vol}_\tau(S_t)$, notice that if a matrix multiplication tensor involves X, Y, Z each of x, y, z-variables, respectively, then its volume is $\sqrt{XYZ}$: indeed, for $\langle n, m, p\rangle$ we have $X = nm$, $Y = mp$, $Z = pn$ and the volume is $\mathrm{Vol}(\langle n, m, p\rangle) = nmp = \sqrt{XYZ}$. Each x-index contains exactly $(\alpha_y + \alpha_z)N$ coordinates equal to 1, and so it corresponds to $q^{(\alpha_y+\alpha_z)N}$ variables. Therefore

$$\mathrm{Vol}_\tau(S_t) = \sqrt{q^{(\alpha_y+\alpha_z)N} I_t q^{(\alpha_x+\alpha_z)N} J_t q^{(\alpha_x+\alpha_y)N} K_t} = q^{\alpha N}\sqrt{I_t J_t K_t}.$$

In total, we obtain the upper bound

$$U_{\rho,N}(\tau_\ell, \tau) \leq q^{(\alpha\rho/3)N} \sum_t (I_t J_t K_t)^{\rho/6}.$$

Let us focus now on the quantity

$$\sigma = \sum_t (I_t J_t K_t)^{\rho/6}.$$

We want to obtain an upper bound on σ. We can assume that $\sum_t I_t = P_x$, $\sum_t J_t = P_y$, $\sum_t K_t = P_z$. Lagrange multipliers show that this quantity is optimized when $I_t^{\rho/6-1}(J_t K_t)^{\rho/6}, J_t^{\rho/6-1}(I_t K_t)^{\rho/6}, K_t^{\rho/6-1}(I_t J_t)^{\rho/6}$ are all constant. Multiplying all these constraints together, we get that $I_t J_t K_t$ is constant (assuming $\rho \neq 2$) and so I_t, J_t, K_t are constant. In order to find the constants, let π be the number of different summands. Then $I_t = P_x/\pi$, $J_t = P_y/\pi$, $K_t = P_z/\pi$. On the other hand, $I_t \leq Q_x$, $J_t \leq Q_y$, $K_t \leq Q_z$, and so $\pi \geq \max(P_x/Q_x, P_y/Q_y, P_z/Q_z) \geq \sqrt[3]{P_x P_y P_z / Q_x Q_y Q_z}$. Therefore

$$\sigma \leq \max_{\pi \geq \sqrt[3]{P_x P_y P_z / Q_x Q_y Q_z}} \pi^{1-\rho/2} (P_x P_y P_z)^{\rho/6}.$$

Since $1 - \rho/2 \leq 0$, we would like π to be as small as possible, and so

$$\sigma \leq (P_x P_y P_z / Q_x Q_y Q_z)^{1/3-\rho/6} (P_x P_y P_z)^{\rho/6} = (P_x P_y P_z)^{1/3} (Q_x Q_y Q_z)^{(\rho-2)/6}.$$

Altogether, we obtain the upper bound

$$U_{\rho,N}(\tau_\ell, \tau) \leq q^{(\alpha\rho/3)N} (P_x P_y P_z)^{1/3} (Q_x Q_y Q_z)^{(\rho-2)/6}.$$

The concavity of the entropy function shows that

$$\begin{aligned}
&\frac{1}{N}\log(P_x P_y P_z)^{1/3} \\
=&\frac{H(\alpha_x+\beta_y+\beta_z, \alpha_y+\alpha_z, \beta_x) + H(\beta_x+\alpha_y+\beta_z, \alpha_x+\alpha_z, \beta_y) + H(\beta_x+\beta_y+\alpha_z, \alpha_x+\alpha_y, \beta_z)}{3} \\
\leq& H(\tfrac{\alpha+2\beta}{3}, \tfrac{2\alpha}{3}, \tfrac{\beta}{3}) = H(\tfrac{2-\alpha}{3}, \tfrac{2\alpha}{3}, \tfrac{1-\alpha}{3}).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\frac{1}{N}\log(Q_x Q_y Q_z)^{1/2} \\
=&\frac{1-\gamma_x}{2} H(\tfrac{\alpha_x+\beta_y+\beta_z-\gamma_x}{1-\gamma_x}, \tfrac{\alpha_y+\alpha_z}{1-\gamma_x}, \tfrac{\beta_x}{1-\gamma_x}) + \frac{1-\gamma_y}{2} H(\tfrac{\beta_x+\alpha_y+\beta_z-\gamma_y}{1-\gamma_y}, \tfrac{\alpha_x+\alpha_z}{1-\gamma_y}, \tfrac{\beta_y}{1-\gamma_y}) + \frac{1-\gamma_z}{2} H(\tfrac{\beta_x+\beta_y+\alpha_z-\gamma_z}{1-\gamma_z}, \tfrac{\alpha_x+\alpha_y}{1-\gamma_z}, \tfrac{\beta_z}{1-\gamma_z}) \\
\leq& H(\tfrac{\alpha+2\beta-1}{2}, \alpha, \tfrac{\beta}{2}) = H(\tfrac{1-\alpha}{2}, \alpha, \tfrac{1-\alpha}{2}).
\end{aligned}$$

Therefore

$$U_{\rho,N}(\tau_\ell, \tau)^{1/N} \leq q^{\alpha\rho/3} \exp_2 H(\tfrac{2-\alpha}{3}, \tfrac{2\alpha}{3}, \tfrac{1-\alpha}{3}) \exp_2[H(\tfrac{1-\alpha}{2}, \alpha, \tfrac{1-\alpha}{2})\tfrac{\rho-2}{3}].$$

The theorem now follows from (1). □

The following table gives the corresponding values of ω (third column). The merging technique cannot produce better bounds on ω. For comparison, the second column presents the values of ω obtained through computing the value.

q	$V_\omega(T_{CW}) = q+2$	$V^*_\omega(T_{CW}) = q+2$
2	2.698562159324317	2.254065971513967
3	2.4739043228079143	2.2725388264017234
4	2.4141016915313966	2.2907590555822686
5	2.393410050907498	2.3078038096573787
6	2.387189908200805	2.323464284487939
7	2.387427468309851	2.3377704174176306
8	2.3908116579971983	2.3508319889917244
9	2.3957738127065547	2.3627788661850917

The best known upper bound on ω, due to Le Gall [7], obtained for $q = 5$ and $\tilde{T}^{\otimes 32}_{CW}$, is $\omega < 2.3728639$. Theorem 4.2 shows that using $q = 5$ and $\tilde{T}^{\otimes N}_{CW}$ for *any* N, the best bound on ω that can possibly be obtained is $\omega < 2.3078039$.

5 Discussion

Our main result shows that the conjecture $\omega = 2$ cannot be proved using the merging technique applied to the tensor T_{CW}. On the other hand, we believe that the technique can be used to improve known bounds on ω. We believe that it is possible that

$$V^*_\rho(T_{CW}) > \limsup_{n\to\infty} V_\rho(\tilde{T}^{\otimes n}_{CW}).$$

The reason is that $V_{\rho,N/n}(\tilde{T}^{\otimes n}_{CW})$ corresponds to a lower bound on $V^*_{\rho,N}(T_{CW})$ in which merging is done in groups of n coordinates at a time, for fixed n; if the merging width n is allowed to vary with N, then a better lower bound on $V^*_{\rho,N}(T_{CW})$ can potentially be obtained.

Our main result gives a limit on the possible upper bounds on ω obtainable for given $q \geq 2$ which deteriorates as q gets smaller. In contrast, for known constructions the best q is $q = 5$ (or $q = 6$ for the construction without merging). This leads us to believe that our upper bound on the extended value is not tight. We leave it as an open question to determine the correct value of $V^*_\rho(T_{CW})$.

A similar issue concerns the generalized value $V_\rho(\tilde{T}^{\otimes n}_{CW})$. The methods of Theorem 2.5 can be used to calculate the value for $n = 1$ and $n = 2$, but already for $n = 4$ there is a gap between the lower bound given by the construction in Stothers [9, 5] and the upper bound given by the method of Cohn et al. [3]; more details appear following Theorem 2.5. We conjecture that the lower bound is tight, but have so far been unable to prove this.

Research in matrix multiplication has proceeded in the past by finding new techniques and new identities (corresponding to upper bounds on ranks or border ranks of tensors). Notwithstanding recent developments, this process seems to have stagnated. In this paper we propose a new technique that could potentially lead to improved bounds on ω. However, we find the most promising research direction to be *finding new identities*. Perhaps a systematic search for new identities can be automated and could lead to significantly improved upper bounds on ω.

References

[1] Dario A. Bini, Mario Capovani, Francesco Romani, and Grazia Lotti. $O(n^{2.7799})$ complexity for $n \times n$ approximate matrix multiplication. *Information Processing Letters*, 8:234–235, 1979.

[2] Peter Bürgisser, Michael Clausen, and M. Amin Shokrollahi. *Algebraic Complexity Theory*. Springer, 1997.

[3] Henry Cohn, Robert Kleinberg, Balázs Szegedy, and Chris Umans. Group-theoritic algorithms for matrix multiplication. In *46th Annual Symposium on Foundations of Computer Science (FOCS 2005)*, 2005.

[4] Dan Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation*, 9:251–280, March 1990.

[5] A. M. Davie and A. J. Stothers. Improved bound for complexity of matrix multiplication. *Proceedings of the Royal Society of Edinburgh*, 143A:351–370, 2013.

[6] Johan Håstad. Tensor rank is NP-complete. *J. Algorithms*, 11(4):644–654, December 1990.

[7] François Le Gall. Powers of tensors and fast matrix multiplication. In *39th International Symposium on Symbolic and Algebraic Computation (ISSAC 2014)*, 2014. arXiv:1401.771.

[8] Arnold Schönhage. Partial and total matrix multiplication. *SIAM J. Comp.*, 10:434–455, 1981.

[9] Andrew James Stothers. *On the complexity of matrix multiplication.* PhD thesis, 2010.

[10] Volker Strassen. Gaussian elimination is not optimal. *Num. Math.*, 13:354–356, 1969.

[11] Volker Strassen. Vermeidung von Divisionen [Avoiding divisions]. *Crelles J. Reine Angew. Math.*, 264:184–202, 1973.

[12] Volker Strassen. Relative bilinear complexity and matrix multiplication. *Crelles J. Reine Angew. Math.*, 375/376:406–443, 1987.

[13] Virginia Vassilevska-Williams. Breaking the Coppersmith–Winograd barrier. In *44th ACM Symposium on Theory of Computing (STOC 2012)*, 2012.