Electronic Colloquium on Computational Complexity, Report No. 298 (2026)

Semidirect sums of tensors and the matrix multiplication exponent

Baitian Li*

Abstract

This note is an exposition of the recent preprints from OpenAI that establish the bound $\omega \leqslant 9/4$ on the matrix multiplication exponent.

We argue that the heart of these preprints is a new technique, which we formulate as the *untwisting lemma* for *semidirect sums* of tensors. This gives a generally applicable method for analyzing tensors. We also discuss further consequences of this method, as well as the relation between the OpenAI proof and previous work.

1 Introduction

Determining the exponent ω of matrix multiplication is one of the foundational questions in algebraic complexity theory. It has been known since Strassen [Str69] that determining ω amounts to studying the asymptotic rank of matrix multiplication tensors. Let $\mathrm{R}(T)$ and $\underline{\mathrm{R}}(T)$ denote the rank and the border rank of a tensor T. The *asymptotic rank* of T is defined by

$$\tilde{\mathrm{R}}(T) \coloneqq \lim_{n\to\infty} \mathrm{R}(T^{\otimes n})^{1/n}.$$

Then $\omega = \log_2 \tilde{\mathrm{R}}(\langle 2,2,2\rangle)$, and more generally $\omega = \log_m \tilde{\mathrm{R}}(\langle m,m,m\rangle)$ for any integer $m \geqslant 2$.

The long quest for the matrix multiplication exponent can be divided into two kinds of efforts: developing new frameworks for bounding the exponent, and discovering instantiations that are effective within such frameworks.

In his foundational work [Str69], Strassen observed that any finite estimate $\mathrm{R}(\langle n,n,n\rangle) \leqslant r$ gives the bound $n^\omega \leqslant r$; together with his discovery that $\mathrm{R}(\langle 2,2,2\rangle) \leqslant 7$, this yields the first nontrivial bound $\omega \leqslant \log_2 7$. It was then observed [BCRL79, Bin80] that border rank bounds for rectangular matrix multiplication also suffice: if $\underline{\mathrm{R}}(\langle E,H,L\rangle) \leqslant r$, then $(EHL)^{\omega/3} \leqslant r$.

The next milestone is Schönhage's celebrated *asymptotic sum inequality* [Sch81], which later became a cornerstone of the study of asymptotic rank. It shows that one can even use border rank bounds for direct sums of matrix multiplication tensors to bound ω:

$$\text{if} \quad \underline{\mathrm{R}}\left(\bigoplus_{i=1}^{q} \langle E_i,H_i,L_i\rangle\right) \leqslant r, \quad \text{then} \quad \sum_{i=1}^{q} (E_iH_iL_i)^{\omega/3} \leqslant r.$$

Schönhage complemented this beautiful inequality with his discovery of the *direct sum identity*

$$\underline{\mathrm{R}}\big(\langle n,1,m\rangle \oplus \langle 1,(n-1)(m-1),1\rangle\big) = nm+1.$$

*Columbia University. bl3052@columbia.edu.

ISSN 1433-8092

This is very surprising: the first summand alone already has border rank nm, yet adding the second summand costs only one extra unit, whereas computing it separately would require $(n-1)(m-1)$.

The next conceptual breakthrough again came from Strassen, who introduced the *laser method* [Str87]. It makes it possible to obtain bounds from tensors that are sums of matrix multiplication tensors which are *not* direct sums. The method was then refined by Coppersmith and Winograd [CW90] to make it more widely applicable, and they applied it to the family of tensors now known as the big Coppersmith–Winograd tensors CW_q. All subsequent improvements over more than three decades [Sto10, DS13, Vas12, Le 14, AV21, DWZ23, VXXZ24, ADV+25, DEK+26] rely on increasingly refined analyses of modified versions of the laser method applied to CW_q. The best published bound along this line of work is $\omega < 2.371177$ [DEK+26], which improves upon the bound $\omega < 2.3755$ of Coppersmith and Winograd only in the third decimal place. Moreover, as shown by Ambainis, Filmus and Le Gall [AFL15], a broad generalization of the laser method applied to CW_q cannot yield any bound better than $\omega < 2.3078$.

Very recently, OpenAI released three preprints [Ope26c, Ope26b, Ope26a] that culminate in the substantial improvement $\omega \leqslant 9/4 = 2.25$. In this exposition, we give a streamlined proof of this result, identify the key technical breakthrough behind it, and discuss its connections with previous work.

To explain these connections, we first go back to the time when Strassen invented the laser method [Str87].

Notation. We follow the notation that has become standard in the matrix multiplication literature over the past decade; we list some of it here. For a detailed treatment of the preliminaries, we refer the reader to [BCS97, Chap. 15] or [Blä13]. A tensor T lives in a tensor space $U \otimes V \otimes W$, and the spaces U, V, W are called the three *modes* of the tensor. Basis vectors of U, V, W are written as x_i, y_j, z_k, respectively. We write $\leqslant$ and $\trianglelefteq$ for restriction and degeneration of tensors, and $\lesssim$ (or $\gtrsim$) for asymptotic restriction. The diagonal tensor is $\langle n \rangle = \sum_{i=1}^{n} x_i y_i z_i$, and the matrix multiplication tensor is $\langle E, H, L \rangle = \sum_{i=1}^{E} \sum_{j=1}^{H} \sum_{k=1}^{L} x_{ij} y_{jk} z_{ki}$.

1.1 The semidirect sum of tensors

Let T_1, T_2 be two tensors whose third modes are the same distinguished space W, that is, $T_1 \in U_1 \otimes V_1 \otimes W$ and $T_2 \in U_2 \otimes V_2 \otimes W$. We call their sum $T_1 + T_2$, viewed in the ambient space $(U_1 \oplus U_2) \otimes (V_1 \oplus V_2) \otimes W$, the *semidirect sum* of T_1 and T_2, and denote it by $T_1 \boxplus_3 T_2$; the subscript 3 indicates that the two tensors share the third mode. When the shared mode is clear from context, we usually drop the subscript.

The Strassen tensor. Semidirect sums of tensors appear frequently in the study of matrix multiplication, although they have not been formally defined. In the paper where Strassen introduced the *laser method* [Str87], the tensor he analyzed is the following one, which he called the null algebra and which is now known as the Strassen tensor:

$$\mathrm{Str}_q = \underbrace{\sum_{i=1}^{q} x_0 y_i z_i}_{T_1} + \underbrace{\sum_{i=1}^{q} x_i y_0 z_i}_{T_2}.$$

This is naturally a semidirect sum.

The tensor Str_q has minimal border rank $q+1$, as the following border decomposition shows:

$$\mathrm{Str}_q = \lim_{\lambda \to 0} \frac{1}{\lambda} \left(\sum_{i=1}^{q} (x_0 + \lambda x_i)(y_0 + \lambda y_i) z_i - x_0 y_0 \sum_{i=1}^{q} z_i \right).$$

An interesting feature of the Strassen tensor is that both of its semidirect summands T_1, T_2 and its "outer structure" are matrix multiplication tensors. More precisely, each of them is an inner product:

- The first summand T_1 is an inner product on the V and W modes, which we write as $\langle 1, 1, q \rangle$.
- The second summand T_2 is an inner product on the U and W modes, which we write as $\langle q, 1, 1 \rangle$.
- Writing $T_1 \in U_1 \otimes V_1 \otimes W$ and $T_2 \in U_2 \otimes V_2 \otimes W$, the *blockwise support* of Str_q is

  $$\mathrm{Str}_q \ \in \ U_1 \otimes V_1 \otimes W \ \oplus \ U_2 \otimes V_2 \otimes W,$$

  which has the shape of the tensor $u_1 v_1 w + u_2 v_2 w = \langle 1, 2, 1 \rangle$.

The "schizophrenia". One might hope to exploit this by swapping the roles of some of the x- and y-variables:

$$\mathrm{Str}_q = \sum_{i=1}^{q} (x_0 y_i + y_0 x_i) z_i.$$

With the variables ordered as above, this looks just like the matrix multiplication tensor $\langle 1, 2, q \rangle$:

$$\sum_{i=1}^{q} (x_0 y_i + y_0 x_i) z_i \quad \text{“} \cong \text{”} \quad \sum_{i=1}^{q} \sum_{j=1}^{2} (x'_{1j} y'_{ji}) z'_{i1} = \langle 1, 2, q \rangle.$$

If this argument were valid, we would obtain something like $\underline{\mathrm{R}}(\langle 1, 2, q \rangle) \leqslant q + 1$, which would in turn give

$$\omega \leqslant 3 \cdot \frac{\log(q+1)}{\log(2q)}. \tag{1.1}$$

Taking $q = 3$, this would give $\omega < 2.322$.

However, the argument is plainly false: the tensor $\langle 1, 2, q \rangle$ is known to have border rank exactly $2q$. The flaw is that permuting variables across different modes destroys the (border) rank decomposition.

What if we return to the semidirect sum formalism instead? We first write Str_q as

$$\mathrm{Str}_q = \langle 1, 1, q \rangle \boxplus \langle q, 1, 1 \rangle,$$

and take its tensor square. By the "binomial theorem", we would obtain

$$\begin{aligned} \mathrm{Str}_q^{\otimes 2} &= \langle 1, 1, q \rangle^{\otimes 2} \boxplus (\langle 1, 1, q \rangle \otimes \langle q, 1, 1 \rangle)^{\boxplus 2} \boxplus \langle q, 1, 1 \rangle^{\otimes 2} \\ &= \langle 1, 1, q^2 \rangle \boxplus \langle q, 1, q \rangle^{\boxplus 2} \boxplus \langle q^2, 1, 1 \rangle. \end{aligned}$$

More generally, the $2n$-th tensor power $\mathrm{Str}_q^{\otimes 2n}$ should contain $\binom{2n}{n}$ semidirect summands of the form $\langle q^n, 1, q^n \rangle$. This suggests the restriction

$$\mathrm{Str}_q^{\otimes 2n} \geqslant \left\langle q^n, \binom{2n}{n}, q^n \right\rangle. \tag{1.2}$$

If this were true, we could use $(q^{2n} \cdot \binom{2n}{n})^{\omega/3} \leqslant (q+1)^{2n}$, which again yields the desired estimate (1.1). The above heuristic leads one to conjecture an "AM–GM type inequality" for semidirect sums of tensors:

$$\langle q,1,1\rangle \boxplus \langle 1,1,q\rangle \gtrsim \text{“}\langle \sqrt{q},2,\sqrt{q}\rangle\text{”},$$

or, more ambitiously, for general tensors, provided a suitable notion of "square root" can be defined:

$$T_1 \boxplus T_2 \gtrsim \text{“}\langle 1,2,1\rangle \otimes \sqrt{T_1} \otimes \sqrt{T_2}\text{”}.$$

However, this argument already breaks down at the very first step. Let us expand the square of the Strassen tensor and examine it carefully:

$$\mathrm{Str}_q^{\otimes 2} = \sum_{i=1}^{q}\sum_{j=1}^{q} x_{00}y_{ij}z_{ij} + \sum_{i=1}^{q}\sum_{j=1}^{q} x_{0j}y_{i0}z_{ij} + \sum_{i=1}^{q}\sum_{j=1}^{q} x_{i0}y_{0j}z_{ij} + \sum_{i=1}^{q}\sum_{j=1}^{q} x_{ij}y_{00}z_{ij}.$$

The "central binomial" part consists of the two middle terms:

$$\mathrm{Str}_q^{\otimes 2} \geqslant \sum_{i=1}^{q}\sum_{j=1}^{q} x_{0j}y_{i0}z_{ij} + \sum_{i=1}^{q}\sum_{j=1}^{q} x_{i0}y_{0j}z_{ij}.$$

In their paper generalizing Strassen's laser method, Coppersmith and Winograd [CW90] reflected on exactly this feature of the Strassen tensor. They described the indices of z as *"schizophrenic"*: in the first summand, i is indexed by y, so it behaves like a column index, while in the second summand, i is indexed by x, so it behaves like a row index. Consequently, these two matrix multiplication tensors cannot be combined into a single larger one!

Let us look at this again through the lens of semidirect sums. The issue is that we wishfully identified $T_1 \otimes T_2 \boxplus T_2 \otimes T_1$ with $T_1 \otimes T_2 \boxplus T_1 \otimes T_2$. Locally, we certainly have $T_1 \otimes T_2 \cong T_2 \otimes T_1$, by permuting the tensor factors within each mode. The problem is that this isomorphism also permutes the factors of the shared third mode, and the semidirect sum is sensitive to exactly this.

For this reason, Strassen did not arrive at the estimate (1.1). Instead, he introduced the laser method to alleviate this "schizophrenia".

The laser method. One first takes the n-th power $\mathrm{Str}_q^{\otimes n}$ and then tensors it with its two cyclic permutations. The resulting tensor has outer structure $\langle 2^n, 2^n, 2^n\rangle$, and each block contains a matrix multiplication tensor of volume q^{3n}. That is, the resulting tensor can be written as

$$T = \sum_{i,j,k=1}^{2^n} T_{ijk}, \qquad T_{ijk} \in U_{ij} \otimes V_{jk} \otimes W_{ki},$$

with respect to direct sum decompositions $U = \bigoplus U_{ij}$, $V = \bigoplus V_{jk}$ and $W = \bigoplus W_{ki}$ of its three modes. The laser method [Str87] shows that one can monomially degenerate the outer structure so that what remains is a direct sum of about 4^n genuine matrix multiplication tensors. Indeed, fix a target integer ℓ, and let λ be a formal parameter. Since

$$(i+j+k-\ell)^2 = \underbrace{(i-\ell)^2 + 2(i-\ell)j}_{U_{ij}} + \underbrace{j^2 + 2jk}_{V_{jk}} + \underbrace{k^2 + 2k(i-\ell)}_{W_{ki}},$$

scaling each block U_{ij}, V_{jk}, W_{ki} by λ to the power of the corresponding underbraced term shows that

$$T \cong \sum_{i,j,k=1}^{2^n} \lambda^{(i+j+k-\ell)^2}\, T_{ijk}.$$

As $(i+j+k-\ell)^2 \geqslant 0$, with equality if and only if $i+j+k=\ell$, letting $\lambda \to 0$ gives the degeneration

$$T \trianglerighteq \bigoplus_{i+j+k=\ell} T_{ijk}.$$

The right-hand side is indeed a direct sum: if $i+j+k=\ell$, then any two of i,j,k determine the third, so distinct summands use disjoint blocks in each mode. Choosing $\ell = 3\cdot 2^{n-1}+1$, the number of summands is exactly $3\cdot 4^{n-1}$. The asymptotic sum inequality then gives

$$3\cdot 4^{n-1}\cdot q^{3n\cdot\omega/3} \leqslant (q+1)^{3n},$$

and letting $n\to\infty$ yields $\omega \leqslant \log_q\big((q+1)^3/4\big)$. Strassen's bound $\omega < 2.479$ follows by plugging in $q=5$.

In other words, instead of letting the outer structure $\langle 2^n, 2^n, 2^n\rangle$ genuinely play the role of a matrix multiplication tensor, which would give

$$\underset{\sim}{\mathrm{R}}(\langle 2,2,2\rangle \otimes \langle q,q,q\rangle) \leqslant (q+1)^3 \implies 2^\omega \cdot q^\omega \leqslant (q+1)^3,$$

the laser method sacrifices the outer structure and changes the game to

$$\underset{\sim}{\mathrm{R}}(\langle 2\cdot 2\rangle \otimes \langle q,q,q\rangle) \leqslant (q+1)^3 \implies 2^2 \cdot q^\omega \leqslant (q+1)^3.$$

Since $\omega \geqslant 2$, the latter can never give a better bound.

The later development of the laser method can be seen as a continued fight against this "schizophrenia". In their celebrated work, Coppersmith and Winograd [CW90] proposed a strengthening of Strassen's laser method, in which the outer structure can be zeroed out to a direct sum even when it is not of matrix multiplication shape. Applying their method to their big tensor CW_q, the "first-order analysis" gives $\omega < 2.388$. They further improved the exponent to $\omega < 2.376$ by analyzing the square $\mathrm{CW}_q^{\otimes 2}$. This is counterintuitive at first glance: both analyses study tensor powers of a fixed tensor, and CW_q and $\mathrm{CW}_q^{\otimes 2}$ have essentially the same family of tensor powers. The subtlety is that, in the analysis of $\mathrm{CW}_q^{\otimes 2}$, they chose a combinatorial outer structure different from the naive square of the first-order one, under which some matrix multiplication tensors merge into larger ones; this advantage is invisible to the first-order analysis. All subsequent improvements [Sto10, DS13, Vas12, Le 14, AV21, DWZ23, VXXZ24, ADV+25, DEK+26] turn out to push this gain further, while balancing it against certain losses in the hierarchical structure of tensor powers.

Nevertheless, whether the "schizophrenia" can be completely "cured" has remained an elusive goal. Even though the naive argument for (1.2) above is false, could the desired restriction nonetheless hold, at least asymptotically? If so, plugging $q=3$ into (1.1) would give $\omega < 2.322$, surpassing the best previous bound, which is based on analyzing CW_5.

The recent preprints from OpenAI [Ope26b, Ope26a] show that the answer is *yes*.

2 The untwisting lemma

In this section we explain the "treatment" of the "schizophrenia" of tensors. For tensors T and S, we say that T *asymptotically restricts* to S, and write $T \gtrsim S$, if $\langle f_n\rangle \otimes T^{\otimes n} \geqslant S^{\otimes n}$ for some $f_n = 2^{o(n)}$.

Theorem 2.1 (Untwisting lemma)**.** *Let $T_1,\dots,T_q$ be tensors with $T_i \in U_i\otimes V_i\otimes W$, each isomorphic to a fixed tensor T, where the isomorphisms are allowed to act nontrivially on the shared third space W. Then*

$$\boxplus_{i=1}^{q} T_i \gtrsim \langle 1,q,1\rangle \otimes T.$$

Although the above lemma requires the tensors T_i to be isomorphic to each other, a standard argument similar to Schönhage's proof of the asymptotic sum inequality [Sch81] boosts it to a version in which the T_i need not be isomorphic. Since we have not yet introduced the language needed to state such a version, we defer it to Section 2.1.

Example 2.2. As a first illustration, the lemma immediately shows that the hypothetical restriction (1.2) we dreamed of earlier does hold *asymptotically*. Recall that the $2n$-th tensor power of $\mathrm{Str}_q = \langle q,1,1\rangle \boxplus \langle 1,1,q\rangle$ is a semidirect sum of 2^{2n} matrix multiplication tensors, exactly $\binom{2n}{n}$ of which are isomorphic to $\langle q^n,1,q^n\rangle$. Restricting to these $\binom{2n}{n}$ summands and applying Theorem 2.1, we obtain

$$\mathrm{Str}_q^{\otimes 2n} \gtrsim \left\langle 1, \binom{2n}{n}, 1 \right\rangle \otimes \langle q^n,1,q^n\rangle = \left\langle q^n, \binom{2n}{n}, q^n \right\rangle,$$

which is the asymptotic version of (1.2). In other words, a single application of Theorem 2.1 to the Strassen tensor already improves Strassen's laser method bound $\omega < 2.479$ (obtained with $q=5$) to $\omega < 2.322$ (obtained with $q=3$)!

Let us now remark on where this lemma comes from. Recently, OpenAI released three preprints improving the upper bound on ω. Chronologically, the first one [Ope26c] improves the previous bound only slightly, via a technically involved higher-order analysis of the laser method.

The second preprint [Ope26b] proves the substantially stronger bound $\omega < 2.258$, and its Lemma 3.3 essentially implies Theorem 2.1. The main test object of the second preprint is the big Coppersmith–Winograd tensor CW_1, which is in fact the structure tensor of the truncated polynomial algebra $\mathbb{F}[x]/(x^3)$.

In the third preprint [Ope26a], the test objects are replaced by the family of polynomial multiplication tensors $\mathbb{F}[x]_{<n} \times \mathbb{F}[x]_{<m} \to \mathbb{F}[x]_{<n+m-1}$, without truncation, which yields the bound $\omega \leqslant 9/4$. We also remark that Proposition 3.1 of the third preprint proves another technical statement which at first glance looks different from Lemma 3.3 of the second preprint; nevertheless, both lead to the same desired result, which we formulate as Theorem 2.1.

We now present the proof of Theorem 2.1, for which we first need a lemma. Let $\langle \mathbb{F}[t]/(t^p)\rangle$ denote the structure tensor of the truncated polynomial algebra, namely

$$\langle \mathbb{F}[t]/(t^p)\rangle = \sum_{\substack{j,k\geqslant 0\\ j+k<p}} x_j y_k z_{j+k}.$$

Lemma 2.3. *Let $T_1,\dots,T_q$ be tensors with $T_i \in U_i \otimes V_i \otimes W$, which need not be isomorphic. If $p \geqslant 2q$, then we have the degeneration*

$$\langle \mathbb{F}[t]/(t^p)\rangle \otimes \boxplus_{i=1}^{q} T_i \trianglerighteq \langle 1,q,1\rangle \otimes \bigoplus_{i=1}^{q} T_i.$$

For readers familiar with the laser method, we remark that the following proof is in fact an "asymmetric" application of Strassen's laser method. Strassen's original laser method [Str87, §7] deals with a *balanced* tight structure, meaning that the numbers of occupied outer blocks in the three modes are the same. Given N blocks in each mode, it shows how to monomially degenerate the structure to $\Omega(N)$ block triples, no two of which share a block in any mode. The asymmetric version starts with an outer block structure with N, N and M blocks in the three modes, where $N \gg M$, and monomially degenerates it to $\Omega(N)$ block triples, any two of which can share a block only in the third mode. This type of asymmetric argument is known as *asymmetric hashing* in the literature [CW90, §9], [DWZ23].

Proof. We first decompose the three mode spaces of the left-hand side into blocks. The first mode is $\mathcal{U} = \mathbb{F}^p \otimes (U_1 \oplus \cdots \oplus U_q)$; expanding $\mathbb{F}^p \otimes U_i = U_i^{\oplus p}$, we write $\mathcal{U}$ as the direct sum of distinguished copies $U_i^{(j)}$ of U_i, where $1 \leqslant i \leqslant q$ and $0 \leqslant j < p$. Similarly, the second mode is the direct sum of copies $V_i^{(k)}$, and the third mode is the direct sum of copies $W^{(\ell)}$, where $0 \leqslant k, \ell < p$. With respect to this block structure, the left-hand side is the sum of distinguished copies $T_i^{(jk)}$ of T_i, one for each i and each pair $j, k \geqslant 0$ with $j + k < p$, where

$$T_i^{(jk)} \in U_i^{(j)} \otimes V_i^{(k)} \otimes W^{(j+k)}.$$

We now construct a monomial degeneration. Write $c_i = i + q - 1$. We multiply the block $T_i^{(jk)}$ by $\lambda^{(j+k-c_i)^2}$, where λ is a formal parameter. This is realized by a monomial degeneration, thanks to the factorization

$$\lambda^{(j+k-c_i)^2} = \underbrace{\lambda^{c_i^2 - 2jc_i}}_{U_i^{(j)}} \cdot \underbrace{\lambda^{-2kc_i}}_{V_i^{(k)}} \cdot \underbrace{\lambda^{\ell^2}}_{W^{(\ell)}}, \qquad \ell = j + k.$$

Note that the exponent on $W^{(\ell)}$ depends only on ℓ and not on i, which is essential since the third mode is shared.

Since $(j + k - c_i)^2$ is a square, it is nonnegative and vanishes only when $j + k = c_i$. Hence, as $\lambda \to 0$, the surviving blocks are exactly those with $j + k = c_i$. This eliminates the variable $k = c_i - j$, so every surviving block $T_i^{(jk)}$ lies in

$$U_i^{(j)} \otimes V_i^{(c_i - j)} \otimes W^{(c_i)}, \qquad 0 \leqslant j \leqslant c_i.$$

Let $T_i^\star$ denote the sum of these surviving blocks. Since $p \geqslant 2q$, we have $c_i \leqslant 2q - 1 < p$, so all $c_i + 1$ such blocks are present. Therefore:

1. The outer structure of $T_i^\star$ is a $(c_i + 1)$-term inner product, hence $T_i^\star \cong \langle 1, q + i, 1\rangle \otimes T_i \geqslant \langle 1, q, 1\rangle \otimes T_i$.
2. The tensors $T_i^\star$ occupy pairwise disjoint subspaces, since $T_i^\star \in U_i^{\oplus p} \otimes V_i^{\oplus p} \otimes W^{(c_i)}$ and the c_i are pairwise distinct.

Consequently,

$$\langle \mathbb{F}[t]/(t^p)\rangle \otimes \boxplus_{i=1}^{q} T_i \;\trianglerighteq\; \bigoplus_{i=1}^{q} T_i^\star \;\geqslant\; \bigoplus_{i=1}^{q} \langle 1, q, 1\rangle \otimes T_i. \qquad \square$$

We now make several remarks on the above proof and its relation to the OpenAI preprints.

Remark 2.4. The third OpenAI preprint essentially carries out the argument of Lemma 2.3 in [Ope26a, Prop. 3.1], whose statement directly combines it with a (border) rank upper bound for $\langle \mathbb{F}[t]/(t^p)\rangle$. We keep these two ingredients separate and use the border rank bound only later, in the proof of Theorem 2.1, so that our lemma isolates the purely combinatorial part of the argument. By comparison, the second preprint [Ope26b, Lem. 3.3] uses the zeroing-out method, which can be seen as the Coppersmith–Winograd version of the above argument.

Remark 2.5. Historically, the laser method of Strassen [Str87] and that of Coppersmith and Winograd [CW90] look superficially different: the former is a monomial degeneration and retains a constant fraction of the coordinates, while the latter is a zeroing out and retains only a subpolynomial fraction of them. They are nevertheless intrinsically related; we give a thorough discussion of this connection in Appendix A.

Proof of Theorem 2.1. Take $p = 2q$ in Lemma 2.3. Since $\langle \mathbb{F}[t]/(t^p)\rangle$ has minimal border rank p, i.e., $\langle p\rangle \trianglerighteq \langle \mathbb{F}[t]/(t^p)\rangle$,[1] we obtain

$$\langle 2q\rangle \otimes \boxplus_{i=1}^{q} T_i \;\trianglerighteq\; \langle \mathbb{F}[t]/(t^{2q})\rangle \otimes \boxplus_{i=1}^{q} T_i \;\trianglerighteq\; \langle 1,q,1\rangle \otimes \bigoplus_{i=1}^{q} T_i.$$

Since every T_i is isomorphic to T, the right-hand side is isomorphic to $\langle 1,q,1\rangle \otimes \langle q\rangle \otimes T$, so

$$\langle 2q\rangle \otimes \boxplus_{i=1}^{q} T_i \;\trianglerighteq\; \langle q\rangle \otimes \langle 1,q,1\rangle \otimes T.$$

Viewing $\langle 2q\rangle$ as $\langle 2\rangle \otimes \langle q\rangle$, we can cancel the factor $\langle q\rangle$ on both sides, at the price of replacing the degeneration by an asymptotic restriction. We remind the reader that this is the *multiplicative cancellation* property, which we prove in Lemma 2.6 below.[2] We get

$$\langle 2\rangle \otimes \boxplus_{i=1}^{q} T_i \gtrsim \langle 1,q,1\rangle \otimes T.$$

It seems that we have lost a factor of 2. However, we can bootstrap the result by the *tensor power trick*. For every $r \geqslant 1$, the r-th tensor power $\left(\boxplus_{i=1}^{q} T_i\right)^{\otimes r}$ is a semidirect sum of q^r isomorphic copies of $T^{\otimes r}$, sharing the third space $W^{\otimes r}$. Applying what we have just proved to it, we obtain

$$\langle 2\rangle \otimes \left(\boxplus_{i=1}^{q} T_i\right)^{\otimes r} \gtrsim \langle 1,q^r,1\rangle \otimes T^{\otimes r} = (\langle 1,q,1\rangle \otimes T)^{\otimes r}.$$

This means that the restriction $\langle f_n\rangle \otimes \left(\boxplus_{i=1}^{q} T_i\right)^{\otimes n} \geqslant (\langle 1,q,1\rangle \otimes T)^{\otimes n}$ holds with $f_n \leqslant 2^{n/r+o(n)}$. Letting $r \to \infty$ shows that the desired asymptotic restriction holds. □

Lemma 2.6 (Multiplicative cancellation)**.** *Let C be a nonzero tensor. If $C \otimes T \trianglerighteq C \otimes S$, then $T \gtrsim S$.*

Proof. We first prove $C \otimes T^{\otimes n} \trianglerighteq C \otimes S^{\otimes n}$ by induction on n. The case $n = 1$ is the assumption. If it holds for $n-1$, then

$$C \otimes T^{\otimes n} = (C \otimes T^{\otimes(n-1)}) \otimes T \trianglerighteq (C \otimes S^{\otimes(n-1)}) \otimes T = (C \otimes T) \otimes S^{\otimes(n-1)} \trianglerighteq C \otimes S^{\otimes n}.$$

Let $r = \underline{\mathrm{R}}(C)$, so that $\langle r\rangle \trianglerighteq C$. Since C is nonzero, we also have $C \geqslant \langle 1\rangle$. Therefore

$$\langle r\rangle \otimes T^{\otimes n} \trianglerighteq C \otimes T^{\otimes n} \trianglerighteq C \otimes S^{\otimes n} \geqslant S^{\otimes n}.$$

This shows that T asymptotically degenerates to S, which we know is equivalent to asymptotic restriction. □

2.1 Functional estimates

A first consequence of the untwisting lemma, which can be seen as a natural generalization of the bound $\omega < 2.322$ shown above, is an estimate, which we call *semi-superadditivity*, for two classes of tensor functionals. The first is the *value*, an auxiliary function introduced by Coppersmith and Winograd [CW90] and used in all later analyses of the laser method. The second comes from the *asymptotic spectrum*, a family of functionals introduced by Strassen [Str88] to systematically study the asymptotic properties of tensors. In Section 3 we will see that both of them can be used to prove $\omega \leqslant 9/4$; the OpenAI preprint [Ope26a] uses the latter.

[1] For instance, a uniform construction is to take $a_k = \lambda^k$: the algebra $\mathbb{F}[t]/(\prod_{k=1}^{p}(t-a_k))$ is isomorphic to $\mathbb{F}^p$ for $\lambda \neq 0$, and it degenerates to $\mathbb{F}[t]/(t^p)$ as $\lambda \to 0$.

[2] More generally, cancellation still holds when both the hypothesis and the conclusion are asymptotic restrictions; see [Str88, Eqn. 1.11] or [WZ26, Lem. 3.7].

Values. For a parameter $\tau \geqslant 2/3$, the value V_τ of a tensor T is defined by

$$V_\tau(T) := \sup\left\{ \left(\sum_{i=1}^{m} (E_i H_i L_i)^\tau\right)^{1/n} : T^{\otimes n} \trianglerighteq \bigoplus_{i=1}^{m} \langle E_i, H_i, L_i\rangle \right\}.$$

The choice of degeneration $\trianglerighteq$ here is merely a convention; it is not hard to see that replacing it by restriction or by asymptotic restriction yields the same function. It is well known that V_τ is superadditive and supermultiplicative [CW90, §8]:

$$V_\tau(T \oplus S) \geqslant V_\tau(T) + V_\tau(S), \qquad V_\tau(T \otimes S) \geqslant V_\tau(T) \cdot V_\tau(S).$$

We now show that V_τ also behaves well under semidirect sums.

Theorem 2.7 (Semi-superadditivity of the value)**.** *For any semidirect sum $T \boxplus S$, we have*

$$V_\tau(T \boxplus S)^{1/\tau} \geqslant V_\tau(T)^{1/\tau} + V_\tau(S)^{1/\tau}.$$

Proof. The proof is a standard concentration argument. For each $0 \leqslant k \leqslant n$, the n-th tensor power $(T \boxplus S)^{\otimes n}$ restricts to a semidirect sum of $\binom{n}{k}$ isomorphic copies of $T^{\otimes k} \otimes S^{\otimes(n-k)}$. Hence by Theorem 2.1, together with $V_\tau(\langle 1, C, 1\rangle) \geqslant C^\tau$ and the supermultiplicativity of V_τ, we obtain

$$V_\tau(T \boxplus S)^n = V_\tau\big((T \boxplus S)^{\otimes n}\big) \geqslant \binom{n}{k}^\tau V_\tau(T)^k \, V_\tau(S)^{n-k}.$$

Taking $1/\tau$-th powers and summing over k, we obtain

$$\begin{aligned}(n+1)\, V_\tau(T \boxplus S)^{n/\tau} &\geqslant \sum_{k=0}^{n} \binom{n}{k} V_\tau(T)^{k/\tau} \, V_\tau(S)^{(n-k)/\tau} \\ &= \big(V_\tau(T)^{1/\tau} + V_\tau(S)^{1/\tau}\big)^n.\end{aligned}$$

Taking n-th roots and letting $n \to \infty$ gives the desired estimate. □

Asymptotic spectrum. A similar inequality holds for spectral points in the sense of Strassen's asymptotic spectrum [Str88, WZ26]. The *asymptotic spectrum of tensors* $\mathcal{X}$ is the set of all functionals ϕ from tensors to nonnegative reals that are

- (monotone) $\phi(S) \leqslant \phi(T)$ whenever $S \leqslant T$;
- (normalized) $\phi(\langle 1\rangle) = 1$;
- (additive) $\phi(S \oplus T) = \phi(S) + \phi(T)$ for all S, T;
- (multiplicative) $\phi(S \otimes T) = \phi(S) \cdot \phi(T)$ for all S, T.

Its elements are called *spectral points*. A basic way to study spectral points is to parameterize them by their behavior on matrix multiplication tensors. It is a standard fact [Str88, Prop. 4.3] that the restriction of any spectral point ϕ to matrix multiplication tensors is determined by three exponents $\theta_1, \theta_2, \theta_3$:

$$\phi(\langle E, H, L\rangle) = E^{\theta_1} H^{\theta_2} L^{\theta_3} \quad \text{for all } E, H, L.$$

We call the set of all exponent vectors $(\theta_1, \theta_2, \theta_3)$ arising in this way the (logarithmic) *spectrum of matrix multiplication*, and denote it by $\Delta_{\mathrm{MM}} \subset \mathbb{R}^3$.

We have the following analogue of Theorem 2.7.

Theorem 2.8 (Semi-superadditivity of spectral points)**.** *Let ϕ be a spectral point and $\varkappa \in \{1,2,3\}$ with $\theta_\varkappa > 0$. Then*

$$\phi(T \boxplus_{\varkappa+1} S)^{1/\theta_\varkappa} \geqslant \phi(T)^{1/\theta_\varkappa} + \phi(S)^{1/\theta_\varkappa},$$

where the index $\varkappa + 1$ is taken modulo 3.

The proof is almost identical; the only difference is that one has to keep track of which mode is shared, which determines the shape of the inner product produced by Theorem 2.1. We omit the details.

2.2 More consequences

We list some further applications of the untwisting lemma.

Support rank. Consider a tensor of the form

$$T = \sum_{i=1}^{E} \sum_{j=1}^{H} \sum_{k=1}^{L} t_{ijk} \cdot x_{ij} y_{jk} z_{ki}$$

with all coefficients $t_{ijk} \neq 0$; in other words, T has the same support as the matrix multiplication tensor $\langle E, H, L\rangle$. The *support rank* $\mathrm{R}_s(\langle E, H, L\rangle)$ is the minimum rank of such a tensor T. Cohn and Umans [CU13] introduced the corresponding *support rank exponent*

$$\omega_s := \inf\{\tau : \mathrm{R}_s(\langle n, n, n\rangle) = O(n^\tau)\}.$$

Clearly $\omega_s \leqslant \omega$, and Cohn and Umans proved the partial converse $\omega \leqslant (3\omega_s - 2)/2$, i.e., $\omega - 2 \leqslant \frac{3}{2}(\omega_s - 2)$. Using the untwisting lemma, we can easily show that the two exponents in fact coincide.

Lemma 2.9. *If T has the same support as $\langle E, H, L\rangle$, then $T \gtrsim \langle E, H, L\rangle$.*

Proof. Grouping the terms of T according to j, we write $T = \sum_{j=1}^{H} T_j$, where

$$T_j = \sum_{i=1}^{E} \sum_{k=1}^{L} t_{ijk} \cdot x_{ij} y_{jk} z_{ki}.$$

For distinct j, the tensors T_j involve disjoint sets of x- and y-variables, while all of them share the z-variables; hence $T = \boxplus_{j=1}^{H} T_j$ is a semidirect sum. Moreover, rescaling $z_{ki} \mapsto t_{ijk}^{-1} z_{ki}$ shows that each T_j is isomorphic to $\langle E, 1, L\rangle$. Note that this isomorphism depends on j and acts nontrivially on the shared third mode, which is exactly the situation allowed by Theorem 2.1. Applying it, we obtain $T \gtrsim \langle 1, H, 1\rangle \otimes \langle E, 1, L\rangle = \langle E, H, L\rangle$. □

Corollary 2.10. $\omega_s = \omega$.

Proof. It suffices to show $\omega \leqslant \omega_s$. For every n, let T be a tensor with the same support as $\langle n, n, n\rangle$ and $\mathrm{R}(T) = \mathrm{R}_s(\langle n, n, n\rangle)$. By Lemma 2.9, $T \gtrsim \langle n, n, n\rangle$, and hence $n^\omega \leqslant \mathrm{R}(T) = \mathrm{R}_s(\langle n, n, n\rangle)$. Therefore $\omega \leqslant \omega_s$. □

Partial matrix multiplication. We keep the notation above, but now allow some of the coefficients t_{ijk} to vanish; we call such a tensor T a *partial and weighted matrix multiplication tensor*, and write $\operatorname{supp} t := \{(i,j,k) : t_{ijk} \neq 0\}$. The argument above generalizes easily to this setting. Such tensors already appeared in the work of Bini, Capovani, Lotti and Romani [BCRL79], where border rank was introduced, and were later studied systematically by Schönhage [Sch81].

More precisely, Schönhage [Sch81, Thm. 4.1] treated the case where T is a *partial* matrix multiplication, i.e., T is obtained from a matrix multiplication tensor by zeroing out some variables in two of its modes. He showed that such a T asymptotically degenerates to a *total* (i.e., genuine) matrix multiplication tensor of volume $|\operatorname{supp} t|$.

The case where $\operatorname{supp} t$ is an arbitrary subset, not necessarily obtained by zeroing out variables, was studied by Vrana [Vra24], who showed that T then asymptotically degenerates to a tensor whose support is that of a matrix multiplication tensor of volume $|\operatorname{supp} t|$. Combined with Lemma 2.9, this shows that T in fact asymptotically degenerates to a genuine matrix multiplication tensor of volume $|\operatorname{supp} t|$. In particular, we obtain the following estimate, for which we give a direct proof via Theorem 2.7.

Theorem 2.11. *Every partial and weighted matrix multiplication tensor T satisfies $V_\tau(T) \geqslant |\operatorname{supp} t|^\tau$.*

Proof. Suppose $\operatorname{supp} t \subseteq [E] \times [H] \times [L]$. We proceed in three steps, peeling off one index at a time.

First, suppose $H = L = 1$. Then

$$T = \sum_{i=1}^{E} t_{i11} \cdot x_{i1} y_{11} z_{1i},$$

which, after discarding the unused variables and rescaling, is isomorphic to $\langle |\operatorname{supp} t|, 1, 1\rangle$. Hence $V_\tau(T) = |\operatorname{supp} t|^\tau$.

Next, suppose only $H = 1$. Grouping the terms according to k, we write

$$T = \sum_{i=1}^{E}\sum_{k=1}^{L} t_{i1k} \cdot x_{i1} y_{1k} z_{ki} = \sum_{k=1}^{L} T_k, \qquad T_k = \sum_{i=1}^{E} t_{i1k} \cdot x_{i1} y_{1k} z_{ki}.$$

The tensors T_k share the x-variables and involve disjoint sets of y- and z-variables, so $T = \boxplus_{k=1}^{L} T_k$ is a semidirect sum sharing the first mode. Each T_k falls into the first case, with support of size $|\operatorname{supp} t \cap ([E] \times \{1\} \times \{k\})|$. Applying Theorem 2.7 repeatedly, we obtain

$$V_\tau(T)^{1/\tau} \geqslant \sum_{k=1}^{L} V_\tau(T_k)^{1/\tau} = \sum_{k=1}^{L} |\operatorname{supp} t \cap ([E] \times \{1\} \times \{k\})| = |\operatorname{supp} t|.$$

Finally, for general H, grouping the terms according to j writes $T = \boxplus_{j=1}^{H} T_j$ as a semidirect sum sharing the third mode, as in the proof of Lemma 2.9, where each T_j falls into the second case. The same computation gives $V_\tau(T)^{1/\tau} \geqslant \sum_{j=1}^{H} V_\tau(T_j)^{1/\tau} \geqslant |\operatorname{supp} t|$. □

2.3 Barriers surpassed by the untwisting lemma

We now have the key tool for proving the new bound $\omega \leqslant 9/4$. Before doing so, we reflect on the effectiveness of this tool and on the barriers it may face. All previous methods for bounding the matrix multiplication exponent can be viewed as consisting of two steps:

(1) start with a tensor T with a sufficiently good asymptotic rank estimate, which usually comes directly from a border rank bound;

(2) determine the most valuable (direct sums of) matrix multiplication tensors to which $T^{\otimes n}$ can be restricted.

The technical crux is how much freedom one has in step (2).

Researchers have proposed several restricted paradigms for step (2) and studied their barriers. The Ambainis–Filmus–Le Gall barrier [AFL15] concerns a variant of the laser method. Since the laser method, as well as the group-theoretic approach of Cohn and Umans [CU03], first fixes a distinguished basis and then analyzes possible zeroings out of the resulting combinatorial structure, Alman and Vassilevska Williams [AV18a, AV18b] formalized this type of approach as the *solar method*. They also considered the more general *galactic method*, which allows monomial degenerations, and the *universal method*, which allows arbitrary degenerations.

We point out that the new estimates obtained via the untwisting lemma are not subject to the Ambainis–Filmus–Le Gall barrier, nor to the known barriers for the solar and galactic methods, since the argument does not fit into the frameworks in which these barriers are proved. At first glance, the untwisting lemma is proved via the laser method itself, but the technical twist is that it introduces a *catalyst*: one first tensors with the truncated polynomial algebra $\langle \mathbb{F}[t]/(t^p)\rangle$. The degeneration obtained from the laser method is combinatorial, but the upper bound on the border rank of $\langle \mathbb{F}[t]/(t^p)\rangle$ is highly non-combinatorial, as it relies on an explicit border rank decomposition. Moreover, this catalyst is effectively used several times, through multiplicative cancellation (Lemma 2.6), which completely scrambles the coefficients of the resulting tensor degeneration.

However, since one still seeks degenerations from tensor powers $T^{\otimes n}$ (even though the new proof does not fit entirely into this framework), the approach remains within the scope of the universal method, and is therefore subject to the slice rank barrier [Alm21], also known as the irreversibility barrier [CVZ21]. For example, even if the untwisting lemma is incorporated into the analysis of the big Coppersmith–Winograd tensors CW_q, one still cannot prove any bound better than $\omega < 2.16805$.

Roughly speaking, suppose that we start with a tensor T with asymptotic rank $\underset{\sim}{\mathrm{R}}(T) = r$. The irreversibility barrier says that if the asymptotic subrank of T is noticeably smaller, say $\underset{\sim}{\mathrm{Q}}(T) \leqslant r^{1-\delta}$, then no analysis based on T can prove a bound better than $\omega < 2/(1-\delta)$. This gives strong barriers for the previously used tensors, since their asymptotic subranks are known to drop, with a uniform bound on δ.

In contrast, the new OpenAI preprint [Ope26a] works with polynomial multiplication tensors. We denote by $P_{p,q}$ the structure tensor of polynomial multiplication $\mathbb{F}[t]_{<p} \times \mathbb{F}[t]_{<q} \to \mathbb{F}[t]_{<p+q-1}$, that is,

$$P_{p,q} = \sum_{i=0}^{p-1} \sum_{j=0}^{q-1} x_i y_j z_{i+j}.$$

It is known that this family of tensors has minimal rank $p+q-1$, and that in the balanced case $p = q$ the tensor $P_{q,q}$ moreover has linear subrank; therefore $\underset{\sim}{\mathrm{Q}}(P_{q,q}) \geqslant \Omega(q)$. In other words, although for each fixed q the irreversibility barrier still rules out proving $\omega = 2$ from $P_{q,q}$ alone, the resulting bounds tend to 2 as $q \to \infty$: for the family $P_{q,q}$ as a whole, we currently know of no barrier towards proving $\omega = 2$! In hindsight, this may explain, at a philosophical level, why the new $\omega \leqslant 9/4$ result is based on this family of tensors.

Compared with the polynomial multiplication tensor $P_{q,q}$, the tensor CW_q seems much less valuable, as it is "hollow" in comparison. The reason that laser method analyses still use the Coppersmith–Winograd tensors is partly that, within those frameworks, there was no good way

to take advantage of the rich middle part of polynomial multiplication tensors. This appears to be no longer the case once semidirect sums are available.

3 Bounds via polynomial multiplication

This section is organized as follows. In the OpenAI preprint [Ope26a], the two key tensor degenerations are intertwined with the inequalities for spectral points that they imply. In Section 3.1, we disentangle them and state them as standalone tensor degenerations. In Section 3.2, we first prove the bound $\omega \leqslant 9/4$ using values, the classical tool of the laser method literature, and then reproduce the proof using spectral points. Finally, in Section 3.3, we take a closer look at the two proofs: beyond the superficial similarity that both reduce to solving the same system of inequalities, we explain the different implications of the bounds for the two types of functionals.

3.1 Constructions

To construct certain tensor degenerations, it is sometimes more convenient to take the viewpoint of bilinear maps instead of tensors. Recall that a bilinear map $T\colon U \times V \to W$ is the same as a tensor in $U^* \otimes V^* \otimes W$. By abuse of notation, we use the same letter for both, and refer to the modes simply as U, V and W. Note that under this identification, a subspace of an input space V corresponds to a quotient of the mode V^*, and vice versa.

Lemma 3.1. *Let $T\colon U \times V \to W$ be a bilinear map, and let $V_1 \subseteq V$ and $W_1 \subseteq W$ be subspaces such that $T(U, V_1) \subseteq W_1$. Then T induces two bilinear maps:*

- *the* subtensor $T_1\colon U \times V_1 \to W_1$, *and*
- *the* quotient tensor $T_2\colon U \times (V/V_1) \to W/W_1$.

Moreover, we have the degeneration $T \trianglerighteq T_1 \boxplus_1 T_2$.

Proof. Choose complements $V = V_1 \oplus V_2$ and $W = W_1 \oplus W_2$. Since $T(U, V_1) \subseteq W_1$, the tensor T has no component in $U \otimes V_1 \otimes W_2$. Scaling V_1 by λ^{-1} and W_1 by λ, the components in $U \otimes V_1 \otimes W_1$ and $U \otimes V_2 \otimes W_2$ are unchanged, while the component in $U \otimes V_2 \otimes W_1$ is multiplied by λ. Letting $\lambda \to 0$ gives a monomial degeneration to $T_1 \boxplus_1 T_2$. □

Proposition 3.2. *For $q \geqslant 2$, we have*

$$\langle 1, 1, 2\rangle \otimes P_{p,q} \trianglerighteq P_{p,q-1} \boxplus_1 P_{p,q+1}.$$

Proof. The left-hand side is the bilinear map $T\colon U \times V \to W$ with

$$U = \mathbb{F}[t]_{<p}, \quad V = \mathbb{F}[t]_{<q}^{\oplus 2}, \quad W = \mathbb{F}[t]_{<p+q-1}^{\oplus 2},$$

given by polynomial multiplication

$$\big(a, (b_1, b_2)\big) \mapsto (ab_1, ab_2),$$

where $a \in \mathbb{F}[t]_{<p}$ and $b_1, b_2 \in \mathbb{F}[t]_{<q}$. Consider the $\mathbb{F}[t]$-linear map

$$\phi\colon \mathbb{F}[t]^{\oplus 2} \to \mathbb{F}[t], \quad (b_1, b_2) \mapsto b_1 + tb_2,$$

and take the subspaces

$$V_1 = (\ker \phi) \cap V, \quad W_1 = (\ker \phi) \cap W,$$

that is, the pairs with $b_1 = -tb_2$. Since ϕ is $\mathbb{F}[t]$-linear, we have $T(U, V_1) \subseteq W_1$. The subtensor $T_1 \colon U \times V_1 \to W_1$ is exactly $P_{p,q-1}$, since $V_1 \cong \mathbb{F}[t]_{<q-1}$ and $W_1 \cong \mathbb{F}[t]_{<p+q-2}$ via $(-tb, b) \mapsto b$. For the quotient tensor $T_2 \colon U \times (V/V_1) \to W/W_1$, the map ϕ induces isomorphisms $V/V_1 \cong \phi(V) = \mathbb{F}[t]_{<q+1}$ and $W/W_1 \cong \phi(W) = \mathbb{F}[t]_{<p+q}$, under which T_2 becomes $P_{p,q+1}$. The claim now follows from Lemma 3.1. □

Let $Q_{p,q}$ denote the tensor obtained from $P_{p,q}$ by swapping the second and third modes, that is,

$$Q_{p,q} = \sum_{i=0}^{p-1} \sum_{j=0}^{q-1} x_i y_{i+j} z_j.$$

Proposition 3.3. *We have*

$$P_{p,3q+p-1} \trianglerighteq (\langle 1, 1, 2 \rangle \otimes P_{p,q}) \boxplus_1 Q_{p,q}.$$

Proof. We split the y-variables into three consecutive blocks

$$Y_1 = \{y_0, \ldots, y_{q-1}\}, \quad Y_2 = \{y_q, \ldots, y_{2q+p-2}\}, \quad Y_3 = \{y_{2q+p-1}, \ldots, y_{3q+p-2}\},$$

of sizes q, $q + p - 1$ and q, respectively. Let $V_1 = \operatorname{span}(Y_1 \sqcup Y_3)$, and let W_1 be the image of $U \times V_1$ under $P_{p,3q+p-1}$. Then W_1 is spanned by the z-variables

$$z_0, \ldots, z_{p+q-2} \quad \text{and} \quad z_{2q+p-1}, \ldots, z_{3q+2p-3}.$$

These two segments do not overlap, so the subtensor $T_1 \colon U \times V_1 \to W_1$ simply multiplies a polynomial in $\mathbb{F}[t]_{<p}$ with each of two independent polynomials in $\mathbb{F}[t]_{<q}$. Hence $T_1 \cong \langle 1, 1, 2 \rangle \otimes P_{p,q}$.

For the quotient tensor $T_2 \colon U \times (V/V_1) \to W/W_1$, the quotient spaces are spanned by Y_2 and by $z_{p+q-1}, \ldots, z_{2q+p-2}$, respectively. Thus

$$T_2 = \sum_{i=0}^{p-1} \sum_{k=0}^{q-1} x_i \, y_{k+q+p-1-i} \, z_{k+p+q-1},$$

and after relabeling $x_i \mapsto x_{p-1-i}$, $y_j \mapsto y_{j-q}$ and $z_\ell \mapsto z_{\ell-p-q+1}$, this is exactly $Q_{p,q}$. The claim now follows from Lemma 3.1. □

3.2 Estimating the exponent

We now have all the tensor degenerations in hand; the remaining task is to analyze how they lead to an estimate of ω. Our overall strategy is still to take advantage of the tensors $P_{p,p}$ of minimal rank $\mathrm{R}(P_{p,p}) = 2p-1$, and to eventually trade these intermediate tensors $P_{p,p}$ for matrix multiplication tensors.

For now, let us ignore the twist that some tensors get their modes permuted in Proposition 3.3, and record the two relations we intuitively have:

- (concavity, Proposition 3.2) $P_{p,q} \to (P_{p,q-1} + P_{p,q+1})/2$;
- (descent, Proposition 3.3) $P_{p,3q+p-1} \to 3P_{p,q}$.

Intuitively, the smaller p or q is, the closer $P_{p,q}$ is to a matrix multiplication tensor; indeed, when $p = 1$ or $q = 1$, it is exactly one. We would therefore like to gradually reduce $P_{p,q}$ to smaller tensors. A natural goal is to turn $P_{p,q+1}$ into several "copies" of $P_{p,q}$, where the copies

are counted not in the direct sum sense, but in the matrix multiplication sense. One can first apply concavity once to obtain $P_{p,q}/2$, at the cost of also producing $P_{p,q+2}/2$. One can then apply concavity again at $(p, q+2)$ to turn the latter into $P_{p,q+1}/4$ and $P_{p,q+3}/4$, and so on. This process does not climb indefinitely: as soon as we reach $P_{p,r}$ with $r = 3q + p - 1$, descent brings it directly back down to $3P_{p,q}$.

To make this involved procedure rigorous, we can use existing tools that reduce it to simple algebraic inequalities. The OpenAI preprint [Ope26a] uses the formalism of the asymptotic spectrum. This is not very surprising, as one motivation of Strassen's theory is to better understand the asymptotic sum inequality [Str88], and there are known examples of bounds obtained via the asymptotic spectrum [AL26]. We point out that one can also use a traditional tool from the laser method literature, namely the notion of *value*, with a slight technical twist. We present both arguments, in Section 3.2.1 and Section 3.2.2 respectively, and then discuss their relation.

3.2.1 Value estimates

We define the *catalytic value* by

$$\overline{V_\tau}(T) := \sup\left\{ \left(\frac{E_1H_1L_1}{E_0H_0L_0}\right)^{\tau/n} : \langle E_0, H_0, L_0\rangle \otimes T^{\otimes n} \trianglerighteq \langle E_1, H_1, L_1\rangle \right\}.$$

Compared with the classical value, there are two technical differences: we allow a catalyst $\langle E_0, H_0, L_0\rangle$ on the left-hand side, but we no longer allow a direct sum on the right-hand side.

Moreover, the parameter τ no longer plays any role: the objective $\left(E_1H_1L_1/(E_0H_0L_0)\right)^{\tau/n}$ is the τ-th power of the same expression with $\tau = 1$, so $\overline{V_\tau}(T) = \overline{V_1}(T)^\tau$. From now on we simply write $\overline{V} = \overline{V_1}$, that is,

$$\overline{V}(T) = \sup\left\{ \left(\frac{E_1H_1L_1}{E_0H_0L_0}\right)^{1/n} : \langle E_0, H_0, L_0\rangle \otimes T^{\otimes n} \trianglerighteq \langle E_1, H_1, L_1\rangle \right\}.$$

The catalytic value is still supermultiplicative, just like the classical value [CW90, §8], but it is no longer obviously superadditive. We do not need superadditivity, since none of the tensor degenerations we use produce direct sums. Moreover, the proof of Theorem 2.7 carries over to the catalytic value, and the catalytic value can also be used to bound ω. We collect these properties below.

Proposition 3.4. *The catalytic value $\overline{V}$ is supermultiplicative and semi-superadditive, that is,*

$$\overline{V}(T \otimes S) \geqslant \overline{V}(T) \cdot \overline{V}(S), \qquad \overline{V}(T \boxplus S) \geqslant \overline{V}(T) + \overline{V}(S),$$

and $\overline{V}(\langle 1\rangle) = 1$. Moreover, it satisfies the following $\langle E, H, L\rangle$-linearity:

$$\overline{V}(\langle E, H, L\rangle \otimes T) = EHL \cdot \overline{V}(T).$$

Finally, we have $\overline{V}(T) \leqslant \underset{\sim}{R}(T)^{3/\omega}$.

Proof. Supermultiplicativity is proved by tensoring degenerations, and semi-superadditivity by the same argument as in the proof of Theorem 2.7; note that the latter only uses the largest term in the binomial expansion, and hence does not use superadditivity. For the normalization, $\overline{V}(\langle 1\rangle) \geqslant 1$ is trivial. Conversely, if $\langle E_0, H_0, L_0\rangle \trianglerighteq \langle E_1, H_1, L_1\rangle$, then comparing the three flattening ranks gives $E_0H_0 \geqslant E_1H_1$, $H_0L_0 \geqslant H_1L_1$ and $L_0E_0 \geqslant L_1E_1$, whose product yields $E_1H_1L_1 \leqslant E_0H_0L_0$. Hence $\overline{V}(\langle 1\rangle) \leqslant 1$.

Next, supermultiplicativity gives $\overline{V}(\langle E,H,L\rangle\otimes T)\geqslant EHL\cdot\overline{V}(T)$. For the reverse inequality, consider any degeneration

$$\langle E_0,H_0,L_0\rangle\otimes(\langle E,H,L\rangle\otimes T)^{\otimes n}\trianglerighteq\langle E_1,H_1,L_1\rangle. \tag{3.1}$$

Its left-hand side equals $\langle E_0E^n,H_0H^n,L_0L^n\rangle\otimes T^{\otimes n}$, so (3.1) is also a catalytic degeneration for T, with the larger catalyst $\langle E_0E^n,H_0H^n,L_0L^n\rangle$. Comparing the two normalizations, the contribution of (3.1) to $\overline{V}(T)$ is exactly $(EHL)^{-1}$ times its contribution to $\overline{V}(\langle E,H,L\rangle\otimes T)$. Taking the supremum gives $\overline{V}(\langle E,H,L\rangle\otimes T)\leqslant EHL\cdot\overline{V}(T)$.

Finally, consider any degeneration

$$\langle E_0,H_0,L_0\rangle\otimes T^{\otimes n}\trianglerighteq\langle E_1,H_1,L_1\rangle.$$

Tensoring it with its two cyclic permutations gives

$$\langle N_0,N_0,N_0\rangle\otimes(T\otimes T^{\sigma}\otimes T^{\sigma^2})^{\otimes n}\trianglerighteq\langle N_1,N_1,N_1\rangle,$$

where $N_0=E_0H_0L_0$, $N_1=E_1H_1L_1$, and σ denotes the cyclic permutation of the modes. Since $\underset{\sim}{R}(\langle N,N,N\rangle)=N^{\omega}$ and $\underset{\sim}{R}(T\otimes T^{\sigma}\otimes T^{\sigma^2})\leqslant\underset{\sim}{R}(T)^3$, taking asymptotic ranks gives $N_1^{\omega}\leqslant N_0^{\omega}\,\underset{\sim}{R}(T)^{3n}$, that is,

$$\left(\frac{E_1H_1L_1}{E_0H_0L_0}\right)^{1/n}\leqslant\underset{\sim}{R}(T)^{3/\omega}.$$

Taking the supremum over all such degenerations gives the desired bound. □

The $\langle E,H,L\rangle$-linearity is the only genuinely new property we need. In Proposition 3.2, we want to estimate the "value" of the left-hand side $\langle 1,1,2\rangle\otimes P_{p,q}$. For the classical value we only know $V_\tau(\langle 1,1,2\rangle\otimes P_{p,q})\geqslant 2^\tau V_\tau(P_{p,q})$, and it is not clear whether equality holds.

Let $\nu_{p,q}=\overline{V}(P_{p,q})$. We now collect the relations satisfied by ν. The catalytic value is invariant under permuting the three modes, since such a permutation maps matrix multiplication tensors to matrix multiplication tensors with the same volume EHL. Since $P_{q,p}$ and $Q_{p,q}$ are both obtained from $P_{p,q}$ by permuting the modes, we have

- (symmetry) $\nu_{p,q}=\nu_{q,p}=\overline{V}(Q_{p,q})$.

By Proposition 3.4, the constructions in the previous subsection imply the following inequalities:

- (Proposition 3.2 $\implies$ concavity) $\nu_{p,q}\geqslant(\nu_{p,q-1}+\nu_{p,q+1})/2$ for $q\geqslant 2$;
- (Proposition 3.3 $\implies$ descent) $\nu_{p,3q+p-1}\geqslant 3\nu_{p,q}$.

Finally, $P_{p,1}\cong\langle p,1,1\rangle$ is itself a matrix multiplication tensor, so by $\langle E,H,L\rangle$-linearity and $\overline{V}(\langle 1\rangle)=1$ we have

- (linear boundary) $\nu_{p,1}=p$.

These relations alone already determine the growth rate of the diagonal, by an elementary calculation.

Lemma 3.5. *Suppose that an array $\nu\colon\mathbb{N}^+\times\mathbb{N}^+\to\mathbb{R}$ satisfies*

- *(symmetry) $\nu_{p,q}=\nu_{q,p}$,*
- *(concavity) $\nu_{p,q}\geqslant(\nu_{p,q-1}+\nu_{p,q+1})/2$ for $q\geqslant 2$,*

- *(descent)* $\nu_{p,3q+p-1} \geqslant 3\nu_{p,q}$, *and*
- *(linear boundary)* $\nu_{p,1} = p$.

Then we have the diagonal estimate $\nu_{p,p} \geqslant p^{4/3}$.

Proof. Write $\Delta\nu_{p,q} = \nu_{p,q+1} - \nu_{p,q}$ for the difference in the second index. By concavity, $\Delta\nu_{p,q}$ is nonincreasing in q.

Applying descent to $P_{p,q}$, we obtain $\nu_{p,3q+p-1} \geqslant 3\nu_{p,q}$, which we rearrange as

$$2\nu_{p,q} \leqslant \nu_{p,3q+p-1} - \nu_{p,q} = \sum_{r=q}^{3q+p-2} \Delta\nu_{p,r} \leqslant (2q+p-1)\,\Delta\nu_{p,q}.$$

This gives

$$(2q+p+1)\,\nu_{p,q} \leqslant (2q+p-1)\,\nu_{p,q+1}.$$

Since $\nu_{p,1} = p > 0$, induction on q shows that $\nu_{p,q} > 0$ for all p, q, so after rearrangement we obtain

$$\frac{\nu_{p,q+1}}{\nu_{p,q}} \geqslant 1 + \frac{2}{2q+p-1}. \tag{3.2}$$

By symmetry, $\nu_{p,p-1} = \nu_{p-1,p}$. Hence, applying (3.2) with $(p,q) \to (p,p-1)$ and with $(p,q) \to (p-1,p-1)$, we get

$$\frac{\nu_{p,p}}{\nu_{p-1,p-1}} = \frac{\nu_{p,p}}{\nu_{p,p-1}} \cdot \frac{\nu_{p-1,p}}{\nu_{p-1,p-1}} \geqslant \left(1+\frac{2}{3p-3}\right)\left(1+\frac{2}{3p-4}\right) = \frac{(3p-1)(3p-2)}{(3p-3)(3p-4)}.$$

Since $\nu_{1,1} = 1$ by the linear boundary condition, induction gives

$$\nu_{p,p} \geqslant \prod_{k=1}^{p-1} \left(1+\frac{2}{3k}\right)\left(1+\frac{2}{3k-1}\right).$$

Finally, we have $1 + \frac{2}{3k-1} \geqslant 1 + \frac{2}{3k}$, and Bernoulli's inequality $(1+x)^{3/2} \geqslant 1 + \frac{3}{2}x$ with $x = \frac{2}{3k}$ gives $\left(1+\frac{2}{3k}\right)^{3/2} \geqslant 1 + \frac{1}{k}$. Therefore,

$$\nu_{p,p} \geqslant \prod_{k=1}^{p-1} \left(1+\frac{2}{3k}\right)^2 \geqslant \prod_{k=1}^{p-1} \left(1+\frac{1}{k}\right)^{4/3} = \left(\prod_{k=1}^{p-1} \frac{k+1}{k}\right)^{4/3} = p^{4/3}. \qquad \square$$

Theorem 3.6. *The matrix multiplication exponent* ω *is at most* $9/4$.

Proof. The tensor $P_{p,p}$ has minimal rank $2p-1$, so by Proposition 3.4 we have $\nu_{p,p} = \overline{V}(P_{p,p}) \leqslant \underset{\sim}{\mathrm{R}}(P_{p,p})^{3/\omega} \leqslant (2p-1)^{3/\omega}$. On the other hand, ν satisfies all the hypotheses of Lemma 3.5, so $\nu_{p,p} \geqslant p^{4/3}$. Hence $p^{4/3} \leqslant (2p-1)^{3/\omega}$ for all $p \geqslant 1$. Taking logarithms and letting $p \to \infty$, we obtain $4/3 \leqslant 3/\omega$, that is, $\omega \leqslant 9/4$. $\square$

Remark 3.7. The original OpenAI preprint [Ope26a] states the bound only over $\mathbb{C}$. However, as we have seen, no step of the argument requires the field to have characteristic zero. The only minor technicality is that the claim that $P_{p,p}$ has rank $2p-1$ requires the ground field $\mathbb{F}$ to have sufficiently many elements. This is harmless: it has been a standard fact since Schönhage [Sch81, Thm. 2.8] that ω depends only on the characteristic of the field, so we may assume without loss of generality that $\mathbb{F}$ is infinite.

In fact, with sufficient care one can deduce the even stronger statement $\omega_{\mathbb{Z}} \leqslant 9/4$, where $\omega_{\mathbb{Z}}$ denotes the exponent of the rank of matrix multiplication tensors over $\mathbb{Z}$ (which in turn bounds the exponent of arithmetic circuits for matrix multiplication with coefficients in $\mathbb{Z}$). To see this, one replaces the rank decomposition of $P_{p,p}$ by a border rank decomposition over $\mathbb{Z}$, similar to the one in the footnote in the proof of Theorem 2.1, and reruns all the machinery of asymptotic rank, together with the above degenerations of polynomial multiplication tensors.

3.2.2 Spectral estimates

Let ϕ be a spectral point. We define the *symmetrization* ϕ^{sym} of ϕ as the geometric mean of the values of ϕ on all mode permutations of T:

$$\phi^{\mathrm{sym}}(T) := \phi\left(\bigotimes_{\sigma\in\mathfrak{S}_3} T^\sigma\right)^{1/6} = \prod_{\sigma\in\mathfrak{S}_3} \phi(T^\sigma)^{1/6},$$

where the second equality holds since spectral points are multiplicative. Since $(T\otimes S)^\sigma = T^\sigma \otimes S^\sigma$, and $T \geqslant S$ implies $T^\sigma \geqslant S^\sigma$, the symmetrization ϕ^{sym} is again multiplicative and monotone under restriction. Moreover, as σ ranges over $\mathfrak{S}_3$, each of E, H, L appears in each position of $\langle E, H, L\rangle^\sigma$ exactly twice, so

$$\phi^{\mathrm{sym}}(\langle E, H, L\rangle) = (EHL)^\tau, \qquad \text{where } \tau = \frac{\theta_1+\theta_2+\theta_3}{3}.$$

It is then not hard to see that ϕ^{sym} shares the other properties of the catalytic value.

Proposition 3.8. *The symmetrization ϕ^{sym} is superadditive and semi-superadditive. More precisely, with $\tau = (\theta_1+\theta_2+\theta_3)/3$, we have*

$$\phi^{\mathrm{sym}}(T\oplus S) \geqslant \phi^{\mathrm{sym}}(T) + \phi^{\mathrm{sym}}(S), \qquad \phi^{\mathrm{sym}}(T\boxplus S)^{1/\tau} \geqslant \phi^{\mathrm{sym}}(T)^{1/\tau} + \phi^{\mathrm{sym}}(S)^{1/\tau},$$

where the semidirect sum may share any of the three modes.

Proof. Write $t_\sigma = \phi(T^\sigma)$ and $s_\sigma = \phi(S^\sigma)$. Since $(T\oplus S)^\sigma = T^\sigma \oplus S^\sigma$ and ϕ is additive, superadditivity amounts to

$$\prod_{\sigma\in\mathfrak{S}_3} (t_\sigma + s_\sigma)^{1/6} \geqslant \prod_{\sigma\in\mathfrak{S}_3} t_\sigma^{1/6} + \prod_{\sigma\in\mathfrak{S}_3} s_\sigma^{1/6},$$

which is Mahler's inequality, i.e., the superadditivity of the geometric mean.

For semi-superadditivity, we run the same concentration argument as in the proof of Theorem 2.7. Suppose that T and S share the $\varkappa$-th mode. For each $0 \leqslant k \leqslant n$, the tensor power $(T\boxplus S)^{\otimes n}$ restricts to a semidirect sum, sharing the $\varkappa$-th mode, of $\binom{n}{k}$ isomorphic copies of $T^{\otimes k}\otimes S^{\otimes(n-k)}$. By Theorem 2.1 (applied after permuting the modes so that the shared mode becomes the third one), this semidirect sum asymptotically restricts to $M\otimes T^{\otimes k}\otimes S^{\otimes(n-k)}$, where M is a matrix multiplication tensor with $EHL = \binom{n}{k}$. Since ϕ^{sym} is multiplicative, monotone, and satisfies $\phi^{\mathrm{sym}}(M) = \binom{n}{k}^\tau$ regardless of which mode is shared, we obtain

$$\phi^{\mathrm{sym}}(T\boxplus S)^n \geqslant \binom{n}{k}^\tau \phi^{\mathrm{sym}}(T)^k\, \phi^{\mathrm{sym}}(S)^{n-k}.$$

Taking $1/\tau$-th powers and summing over k, we get

$$(n+1)\,\phi^{\mathrm{sym}}(T\boxplus S)^{n/\tau} \geqslant \sum_{k=0}^{n} \binom{n}{k} \phi^{\mathrm{sym}}(T)^{k/\tau}\, \phi^{\mathrm{sym}}(S)^{(n-k)/\tau} = \left(\phi^{\mathrm{sym}}(T)^{1/\tau} + \phi^{\mathrm{sym}}(S)^{1/\tau}\right)^n.$$

Taking n-th roots and letting $n\to\infty$ gives the claim. □

We now sketch how to reprove Theorem 3.6 using spectral points. First, it is not hard to see that setting

$$\nu_{p,q} = \phi^{\mathrm{sym}}(P_{p,q})^{1/\tau}$$

again yields an array ν satisfying the hypotheses of Lemma 3.5: symmetry holds since ϕ^{sym} is invariant under permuting the modes, concavity and descent follow from Propositions 3.2 and 3.3 together with Proposition 3.8, and the linear boundary condition follows from $\phi^{\mathrm{sym}}(\langle p,1,1\rangle) = p^{\tau}$. Hence we again conclude that $\phi^{\mathrm{sym}}(P_{p,p}) \geqslant p^{4\tau/3}$.

On the other hand, we recall Strassen's *duality theorem* for asymptotic rank (cf. [Str88, Thm. 3.2], [WZ26, §3]).

Theorem 3.9 (Strassen duality)**.** *For every tensor T, we have*

$$\underset{\sim}{\mathrm{R}}(T) = \max_{\phi \in \mathcal{X}} \phi(T).$$

In particular, since $\phi(\langle n,n,n\rangle) = n^{\theta_1+\theta_2+\theta_3}$, applying Theorem 3.9 to $\langle n,n,n\rangle$ shows that ω equals the maximum of $\theta_1+\theta_2+\theta_3 = 3\tau$ over all spectral points, and this maximum is attained.

Second proof of Theorem 3.6. Let ϕ be a spectral point attaining $\theta_1+\theta_2+\theta_3 = \omega$, so that $\tau = \omega/3$. Since every spectral point is bounded by the rank, and $\mathrm{R}(P_{p,p}^{\sigma}) = \mathrm{R}(P_{p,p})$ for every $\sigma \in \mathfrak{S}_3$, evaluating ϕ^{sym} on $P_{p,p}$ gives

$$p^{4\tau/3} \leqslant \phi^{\mathrm{sym}}(P_{p,p}) \leqslant \mathrm{R}(P_{p,p}) \leqslant 2p-1.$$

Letting $p \to \infty$, we get $4\tau/3 \leqslant 1$, that is, $\omega = 3\tau \leqslant 9/4$. □

3.3 A closer look

As we have seen, both types of functionals lead to the same system of inequalities for the array $\nu_{p,q}$, and hence to the diagonal estimates

$$\overline{V}(P_{p,p}) \geqslant p^{4/3}, \qquad \phi^{\mathrm{sym}}(P_{p,p}) \geqslant p^{4\tau/3}.$$

We point out that each of these inequalities in fact shows that $P_{p,p}$, suitably symmetrized, asymptotically restricts to certain matrix multiplication tensors.

Spectral estimates. We begin by explaining what the estimate

$$\phi^{\mathrm{sym}}(P_{p,p}) \geqslant p^{4\tau/3}$$

says. Unpacking the definition, we write the symmetrization of a tensor as

$$T^{\mathrm{sym}} = \bigotimes_{\sigma \in \mathfrak{S}_3} T^{\sigma},$$

so that $\phi^{\mathrm{sym}}(T) = \phi(T^{\mathrm{sym}})^{1/6}$, and the estimate reads $\phi(P_{p,p}^{\mathrm{sym}}) \geqslant p^{8\tau}$. Recalling that $\tau = (\theta_1+\theta_2+\theta_3)/3$ and taking cubes, we obtain

$$\phi\big((P_{p,p}^{\mathrm{sym}})^{\otimes 3}\big) \geqslant p^{8(\theta_1+\theta_2+\theta_3)} = \phi(\langle p^8, p^8, p^8\rangle).$$

Since this holds for every spectral point ϕ, it is in fact a statement about asymptotic restriction. This is the full version of Strassen duality (cf. [Str88, Thm. 3.2], [WZ26, §3]), of which Theorem 3.9 only states the consequence for asymptotic rank.

Theorem 3.10 (Strassen duality for asymptotic restriction)**.** *For tensors T and S, we have $T \gtrsim S$ if and only if*

$$\phi(T) \geqslant \phi(S) \quad \text{for all } \phi \in \mathcal{X}.$$

Hence Theorem 3.10 gives the asymptotic restriction

$$(P_{p,p}^{\mathrm{sym}})^{\otimes 3} \gtrsim \langle p^8, p^8, p^8 \rangle.$$

In other words, $P_{p,p}^{\mathrm{sym}}$ asymptotically restricts to a square matrix multiplication tensor of volume p^8.

Value estimates. Does this say anything about $P_{p,p}$ itself, before symmetrization? For instance, does $P_{p,p}$ asymptotically restrict to a matrix multiplication tensor of volume $p^{4/3}$? This is not clear from the analysis above. At first glance, the catalytic value estimate

$$\overline{V}(P_{p,p}) \geqslant p^{4/3}$$

does not directly help either: literally, it only provides catalytic degenerations $\langle E_0, H_0, L_0 \rangle \otimes P_{p,p}^{\otimes n} \trianglerighteq \langle E_1, H_1, L_1 \rangle$ whose *relative* volume $E_1 H_1 L_1 / (E_0 H_0 L_0)$ grows like $p^{(4/3 - o(1))n}$. However, we now show that the catalyst, although it seems necessary at first glance, can in fact be removed.

Proposition 3.11. *Let $T \in \mathbb{F}^r \otimes \mathbb{F}^r \otimes \mathbb{F}^r$ be a tensor. If $\overline{V}(T) > r$, then the catalyst can be removed, in the sense that*

$$\overline{V}(T) = \sup\Big\{ (EHL)^{1/n} : T^{\otimes n} \gtrsim \langle E, H, L \rangle \Big\}.$$

Proof. We begin by analyzing an arbitrary catalytic degeneration

$$\langle E_0, H_0, L_0 \rangle \otimes T^{\otimes n} \trianglerighteq \langle E_1, H_1, L_1 \rangle.$$

In principle, we can eliminate the catalyst by the following argument. First, we continue the degeneration by a restriction:

$$\langle E_0, H_0, L_0 \rangle \otimes T^{\otimes n} \trianglerighteq \langle E_1, H_1, L_1 \rangle \geqslant \langle E_0, H_0, L_0 \rangle \otimes \langle \lfloor E_1/E_0 \rfloor, \lfloor H_1/H_0 \rfloor, \lfloor L_1/L_0 \rfloor \rangle.$$

By multiplicative cancellation (Lemma 2.6), we can remove the catalyst from both sides and obtain

$$T^{\otimes n} \gtrsim \langle \lfloor E_1/E_0 \rfloor, \lfloor H_1/H_0 \rfloor, \lfloor L_1/L_0 \rfloor \rangle.$$

However, this can be useless: one of the side lengths of this matrix multiplication tensor may be zero. We show that this cannot happen when $E_1 H_1 L_1 / (E_0 H_0 L_0) > r^n$, which, by the assumption $\overline{V}(T) > r$, is the case for degenerations whose contributions are close to $\overline{V}(T)$. To see this, we compare the flattening ranks of the first mode on both sides of the degeneration. Since the flattening ranks of $\langle E, H, L \rangle$ are EH, HL and LE, we get

$$E_0 H_0 \cdot r^n \geqslant E_1 H_1,$$

that is,

$$\frac{E_1 H_1}{E_0 H_0} \leqslant r^n < \frac{E_1 H_1 L_1}{E_0 H_0 L_0},$$

and we conclude that $L_1 > L_0$. Comparing the flattening ranks of the other two modes, we also conclude that $E_1 > E_0$ and $H_1 > H_0$.

The rounding still loses a constant factor, but this loss can be amplified away by taking the k-th tensor power of the degeneration before removing the catalyst, which gives

$$T^{\otimes nk} \gtrsim \langle \lfloor (E_1/E_0)^k \rfloor, \lfloor (H_1/H_0)^k \rfloor, \lfloor (L_1/L_0)^k \rfloor \rangle.$$

Since the three ratios exceed 1, the volume of the right-hand side is at least $\frac{1}{8}(E_1H_1L_1/E_0H_0L_0)^k$, and its (nk)-th root tends to $(E_1H_1L_1/E_0H_0L_0)^{1/n}$ as $k \to \infty$. Taking a family of catalytic degenerations whose contributions $(E_1H_1L_1/E_0H_0L_0)^{1/n}$ tend to $\overline{V}(T)$, we conclude. □

Since the flattening ranks of $P_{p,p}$ are p, p and $2p-1$, and $p^{4/3} > 2p-1$ for $p \geqslant 7$, Proposition 3.11 applies to $P_{p,p}$ for all $p \geqslant 7$. Hence we can conclude that $P_{p,p}$ asymptotically restricts to matrix multiplication of volume $p^{4/3}$, in the sense that $P_{p,p}^{\otimes n} \gtrsim \langle E, H, L \rangle$ for some E, H, L with $EHL \geqslant p^{(4/3-o(1))n}$.

Rectangular matrices. However, since the value does not distinguish the three side lengths E, H, L, we do not know exactly to which shape of matrix multiplication tensor $P_{p,p}$ asymptotically restricts. In principle, this can be read off from the spectral points. For reals $E, H, L \geqslant 1$, it essentially follows from Theorem 3.10 (applied to tensor powers, up to rounding) that $T^{\otimes n} \gtrsim \langle E^{n-o(n)}, H^{n-o(n)}, L^{n-o(n)} \rangle$ holds if and only if

$$\phi(T) \geqslant E^{\theta_1} H^{\theta_2} L^{\theta_3} \quad \text{for all } \phi \in \mathcal{X}.$$

To exploit this, we have to go back to the spectral values of the polynomial multiplication tensors without symmetrization. In principle, we should then keep track of three arrays rather than just one, corresponding to the three choices of which mode of the polynomial multiplication tensor serves as the output mode. The degenerations we constructed then lead to a more complicated family of inequalities relating these three arrays, since they relate a polynomial multiplication tensor to its mode permutations: for instance, the degeneration of Proposition 3.3 involves both $P_{p,q}$ and $Q_{p,q}$, the latter being $P_{p,q}$ with its last two modes swapped.

From this perspective, symmetrizing the spectral points is essentially a trick to simplify the analysis: it replaces the three exponents $\theta_1, \theta_2, \theta_3$ by their average $\tau = (\theta_1 + \theta_2 + \theta_3)/3$, so that we no longer need to track them separately. This trick works as long as we only care about the square exponent ω. However, if we want bounds on the rectangular exponents $\omega(a, b, c)$, symmetrization loses too much information.

This asymmetric analysis has been carried out by Uznański [Uzn26].[3] Its main result is the following estimate for Δ_{MM}. The reader may also consult the recent exposition of Amanov, van den Berg and Zuiddam [AvdBZ26], which derives this estimate and its consequence Corollary 3.13 below in the language of the asymptotic spectrum [AvdBZ26, §§4.5–4.6], and moreover shows that the current ingredients cannot improve these bounds.

Theorem 3.12 ([Uzn26, Prop. 5.2]; [AvdBZ26, Prop. 4.14])**.** *Every* $(\theta_1, \theta_2, \theta_3) \in \Delta_{\mathrm{MM}}$ *satisfies*

$$\theta_1 + \theta_2 + \theta_3 \leqslant 2 + 2\sqrt{(1-\theta_1)(1-\theta_2)(1-\theta_3)}.$$

Together with the lower bound $\theta_1 + \theta_2 + \theta_3 \geqslant 2$, which follows from the subrank of matrix multiplication [Str87, Thm. 6.6], Theorem 3.12 confines Δ_{MM} to the thin region shown in Figure 1.

[3] The note [Uzn26] was written by an AI model, and its terminology differs from the one commonly used in the literature (and in this note). We have restated its two main results in the standard language.

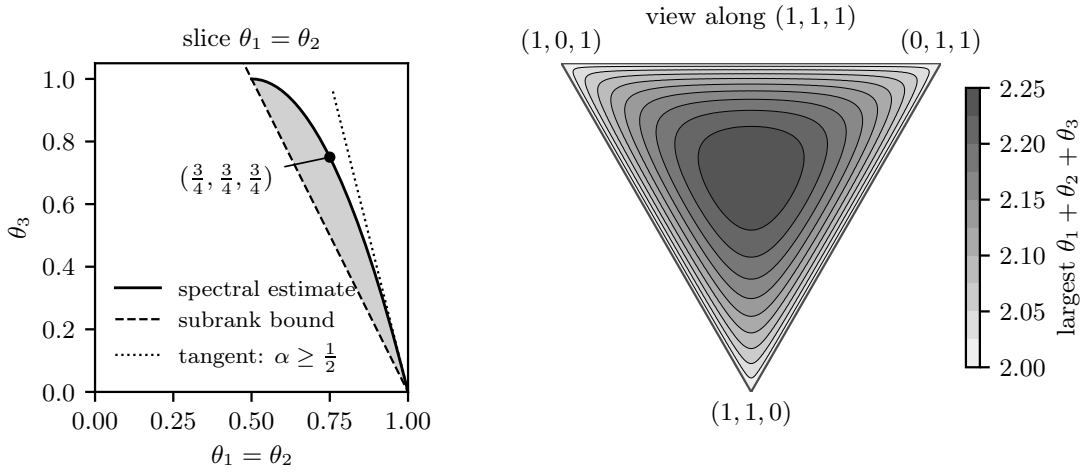


Figure 1: The region containing Δ_{MM}.

Using the basic characterization

$$\omega(a,b,c) = \max_{\theta \in \Delta_{\mathrm{MM}}} (a\theta_1 + b\theta_2 + c\theta_3),$$

which follows from Theorem 3.9 applied to $\langle n^a, n^b, n^c \rangle$, one obtains the following bounds for rectangular matrix multiplication.

Corollary 3.13 ([Uzn26, Thm. 1.1]; [AvdBZ26, Thm. 1.2]). *Let $a \geqslant b \geqslant c \geqslant 0$ with $b > 0$.*

(i) If $c \leqslant ab/(a+b)$, then $\omega(a,b,c) = a + b$.

(ii) Otherwise, $\omega(a,b,c) \leqslant \dfrac{(ab+bc+ca)^2}{4abc}$.

For $a = b = c = 1$, case (ii) gives back $\omega \leqslant 9/4$. For $\omega(1,k,1)$, case (i) applies exactly when $k \leqslant 1/2$, and otherwise case (ii) gives $\omega(1,k,1) \leqslant 1 + k + \frac{1}{4k}$; in particular, the dual exponent satisfies $\alpha \geqslant 1/2$ and we also have $\omega(1,2,1) \leqslant 25/8$.

Background and acknowledgments

Background. In the spring of 2026, while studying Strassen's support functionals [Str91] and the laser method, the author came up with the concept of semidirect sums, together with a conjecture formulated as Theorem 2.1 and Theorem 2.8 of this note (it is not hard to see that these two statements are equivalent). One motivation for this conjecture is that it would improve various estimates in laser method analyses, e.g., it would give the hypothetical bound $\omega < 2.322$ discussed in Section 1.1.

After testing the inequality of Theorem 2.8 on the support functionals, which by [SDW26] are spectral points (and in fact the only explicitly known ones), the author found that it holds for them, and started to believe that it might hold in general. However, a proof remained out of reach despite several attempts, computational experiments, and consultations with AI models.

It therefore came as a great surprise to the author that the technical lemmata of the recent OpenAI preprints [Ope26b, Ope26a] settle exactly this conjecture.

In the author's view, although the new bounds on ω are significant and exciting, these lemmata are the main conceptual breakthrough of the OpenAI preprints: they provide new general methods for analyzing tensors, while the specific bounds can be seen as the outcome of a careful search for candidate tensors to which these methods apply. Since the attribution in the original preprints is poor, the author was motivated to write this exposition, in order to point out the techniques behind the original proofs of these lemmata, to discuss their relation to previous work, and to share his own thoughts about semidirect sums (most of Section 1 and part of Section 2), which had crystallized before he learned of the proof of his conjecture.

Concurrent writings. After the OpenAI release, several groups of people have shared their own understanding of the new bound, each with a focus different from ours. The exposition of Amanov, van den Berg and Zuiddam [AvdBZ26] thoroughly works out the details of the proof, and shows that the analyses of the currently known degenerations of polynomial multiplication tensors are *optimal* [AvdBZ26, §5]. The notes of Pass [Pas26] and Shpilka [Shp26] study the constructiveness of the $\omega \leqslant 9/4$ proof, and give quantitative bounds on the $o(1)$ term behind the $n^{9/4+o(1)}$ conclusion.

Acknowledgments. The author would like to thank Josh Alman and Kevin Pratt for inspiring discussions, for their continuous encouragement to think about semidirect sums and to write this exposition, and for kindly declining coauthorship. He also thanks them and J. M. Landsberg for valuable comments on an earlier draft of this note.

AI use. Claude Code was used for language editing, proofreading, checking computations, and producing Figure 1. The author takes full responsibility for the correctness and presentation of the content.

References

[ADV+25] Josh Alman, Ran Duan, Virginia Vassilevska Williams, Yinzhan Xu, Zixuan Xu, and Renfei Zhou. More asymmetry yields faster matrix multiplication. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2005–2039, 2025. doi:10.1137/1.9781611978322.63. ↑2, ↑5

[AFL15] Andris Ambainis, Yuval Filmus, and François Le Gall. Fast matrix multiplication: Limitations of the Coppersmith-Winograd method. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 585–593, 2015. doi:10.1145/2746539.2746554. ↑2, ↑12

[AL26] Josh Alman and Baitian Li. Asymptotic rank speedup theorems, revisited. In *Proceedings of the 41st Computational Complexity Conference (CCC)*, volume 383 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 36:1–36:42, 2026. doi:10.4230/LIPIcs.CCC.2026.36. ↑15

[Alm21] Josh Alman. Limits on the universal method for matrix multiplication. *Theory Comput.*, 17:Paper No. 1, 2021. doi:10.4086/toc.2021.v017a001. ↑12

[AV18a] 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)*, volume 94 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 25:1–25:15, 2018. doi:10.4230/LIPIcs.ITCS.2018.25. ↑12, ↑29

[AV18b] 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)*, pages 580–591, 2018. doi:10.1109/FOCS.2018.00061. ↑12, ↑29

[AV21] Josh Alman and Virginia Vassilevska Williams. A refined laser method and faster matrix multiplication. In *Proceedings of the 2021 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 522–539, 2021. doi:10.1137/1.9781611976465.32. ↑2, ↑5

[AvdBZ26] Alimzhan Amanov, Maxim van den Berg, and Jeroen Zuiddam. Multiplying matrices via polynomials: Exposition and barriers, 2026. arXiv:2610.30054. ↑21, ↑22, ↑23

[BCRL79] Dario Bini, Milvio Capovani, Francesco Romani, and Grazia Lotti. $O(n^{2.7799})$ complexity for $n \times n$ approximate matrix multiplication. *Inf. Process. Lett.*, 8(5):234–235, 1979. doi:10.1016/0020-0190(79)90113-3. ↑1, ↑11

[BCS97] Peter Bürgisser, Michael Clausen, and M. Amin Shokrollahi. *Algebraic Complexity Theory*, volume 315 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 1997. doi:10.1007/978-3-662-03338-8. ↑2, ↑28

[Beh46] Felix A. Behrend. On sets of integers which contain no three terms in arithmetical progression. *Proc. Natl. Acad. Sci. USA*, 32(12):331–332, 1946. doi:10.1073/pnas.32.12.331. ↑29, ↑31

[Bin80] Dario Bini. Relations between exact and approximate bilinear algorithms. Applications. *Calcolo*, 17(1):87–97, 1980. doi:10.1007/BF02575865. ↑1

[Blä13] Markus Bläser. *Fast Matrix Multiplication*. Number 5 in Theory of Computing Graduate Surveys. Theory of Computing Library, 2013. doi:10.4086/toc.gs.2013.005. ↑2, ↑28

[BS23] Thomas F. Bloom and Olof Sisask. The Kelley–Meka bounds for sets free of three-term arithmetic progressions. *Essent. Number Theory*, 2(1):15–44, 2023. doi:10.2140/ent.2023.2.15. ↑29

[CFTZ22] Matthias Christandl, Omar Fawzi, Hoang Ta, and Jeroen Zuiddam. Larger corner-free sets from combinatorial degenerations. In *Proceedings of the 13th Innovations in Theoretical Computer Science Conference (ITCS)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 48:1–48:20, 2022. doi:10.4230/LIPIcs.ITCS.2022.48. ↑27

[CU03] Henry Cohn and Christopher Umans. A group-theoretic approach to fast matrix multiplication. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 438–449, 2003. doi:10.1109/SFCS.2003.1238217. ↑12

[CU13] Henry Cohn and Christopher Umans. Fast matrix multiplication using coherent configurations. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1074–1087, 2013. doi:10.1137/1.9781611973105.77. ↑10

[CVZ21] Matthias Christandl, Péter Vrana, and Jeroen Zuiddam. Barriers for fast matrix multiplication from irreversibility. *Theory Comput.*, 17:Paper No. 2, 2021. doi:10.4086/toc.2021.v017a002. ↑12

[CVZ23] Matthias Christandl, Péter Vrana, and Jeroen Zuiddam. Universal points in the asymptotic spectrum of tensors. *J. Amer. Math. Soc.*, 36(1):31–79, 2023. doi:10.1090/jams/996. ↑29

[CW90] Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. *J. Symb. Comput.*, 9(3):251–280, 1990. doi:10.1016/S0747-7171(08)80013-2. ↑2, ↑4, ↑5, ↑6, ↑7, ↑8, ↑9, ↑15, ↑27, ↑28

[DEK+26] Emilien Dupont, Marvin Eisenberger, Borislav Kozlovskii, Abbas Mehrabian, Francisco J. R. Ruiz, Abigail See, Renfei Zhou, Josh Alman, Virginia Vassilevska Williams, and Matej Balog. Improving the matrix multiplication exponent with modern optimization and AlphaEvolve, 2026. arXiv:2608.16884. ↑2, ↑5

[DS13] Alexander M. Davie and Andrew James Stothers. Improved bound for complexity of matrix multiplication. *Proc. Roy. Soc. Edinburgh Sect. A*, 143(2):351–369, 2013. doi:10.1017/S0308210511001648. ↑2, ↑5

[DWZ23] Ran Duan, Hongxun Wu, and Renfei Zhou. Faster matrix multiplication via asymmetric hashing. In *Proceedings of the 64th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 2129–2138, 2023. doi:10.1109/FOCS57990.2023.00130. ↑2, ↑5, ↑6

[KM23] Zander Kelley and Raghu Meka. Strong bounds for 3-progressions. In *Proceedings of the 64th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 933–973, 2023. doi:10.1109/FOCS57990.2023.00059. ↑29

[Le 14] François Le Gall. Powers of tensors and fast matrix multiplication. In *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation (ISSAC)*, pages 296–303, 2014. doi:10.1145/2608628.2608664. ↑2, ↑5

[Ope26a] OpenAI. An Upper Bound of 9/4 for the Matrix Multiplication Exponent. OpenAI Math Release preprint OAI:Matrix-Multiplication-Nine-Fourths-October-2-2026, 2026. ↑2, ↑5, ↑6, ↑7, ↑8, ↑12, ↑13, ↑15, ↑17, ↑23

[Ope26b] OpenAI. Complex Matrix Multiplication Below 2.258 and Rectangular Bounds. OpenAI Math Release preprint OAI:Complex-Matrix-Multiplication-Below-2.258-and-Rectangular-Bounds-September-24-2026, 2026. ↑2, ↑5, ↑6, ↑7, ↑23

[Ope26c] OpenAI. Staggered extraction for exact matrix multiplication over every field. OpenAI Math Release preprint OAI:Staggered-extraction-for-exact-matrix-multiplication-over-every-field-September-24-2026, 2026. ↑2, ↑6

[Pas26] Rafael Pass. Matrix multiplication from polynomial packings: A constructive and elementary proof of $\omega \leqslant 9/4$, 2026. arXiv:2610.28627. ↑23

[Rag26] Rushil Raghavan. Improved bounds for 3-progressions, 2026. arXiv:2603.27045. ↑29

[Rot53] Klaus F. Roth. On certain sets of integers. *J. Lond. Math. Soc.*, s1-28(1):104–109, 1953. doi:10.1112/jlms/s1-28.1.104. ↑29

[Sch81] A. Schönhage. Partial and total matrix multiplication. *SIAM J. Comput.*, 10(3):434–455, 1981. doi:10.1137/0210032. ↑1, ↑6, ↑11, ↑17

[SDW26] Keiya Sakabe, M. Levent Doğan, and Michael Walter. Strassen's support functionals coincide with the quantum functionals. In *Proceedings of the 67th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2026. To appear. arXiv:2601.21553. ↑22

[Shp26] Amir Shpilka. Exposition of OpenAI's matrix multiplication in exponent 9/4 (and an explicit algorithm). Electronic Colloquium on Computational Complexity (ECCC), TR26-281, 2026. URL: https://eccc.weizmann.ac.il/report/2026/281/. ↑23

[Sto10] Andrew James Stothers. *On the Complexity of Matrix Multiplication*. PhD thesis, University of Edinburgh, 2010. URL: http://hdl.handle.net/1842/4734. ↑2, ↑5

[Str69] Volker Strassen. Gaussian elimination is not optimal. *Numer. Math.*, 13:354–356, 1969. doi:10.1007/BF02165411. ↑1

[Str87] V. Strassen. Relative bilinear complexity and matrix multiplication. *J. Reine Angew. Math.*, 375/376:406–443, 1987. doi:10.1515/crll.1987.375-376.406. ↑2, ↑4, ↑6, ↑7, ↑21, ↑27, ↑28

[Str88] V. Strassen. The asymptotic spectrum of tensors. *J. Reine Angew. Math.*, 384:102–152, 1988. doi:10.1515/crll.1988.384.102. ↑8, ↑9, ↑15, ↑19

[Str91] V. Strassen. Degeneration and complexity of bilinear maps: Some asymptotic spectra. *J. Reine Angew. Math.*, 413:127–180, 1991. doi:10.1515/crll.1991.413.127. ↑22

[Uzn26] Przemysław Uznański. Rectangular matrix multiplication from polynomial products, 2026. arXiv:2610.10494v2. ↑21, ↑22

[Vas12] Virginia Vassilevska Williams. Multiplying matrices faster than Coppersmith-Winograd. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC)*, pages 887–898, 2012. doi:10.1145/2213977.2214056. ↑2, ↑5

[Vra24] Péter Vrana. Partial and weighted matrix multiplication, 2024. arXiv:2408.15728. ↑11

[VXXZ24] Virginia Vassilevska Williams, Yinzhan Xu, Zixuan Xu, and Renfei Zhou. New bounds for matrix multiplication: From alpha to omega. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3792–3835, 2024. doi:10.1137/1.9781611977912.134. ↑2, ↑5

[WZ26] Avi Wigderson and Jeroen Zuiddam. Asymptotic spectra: Theory, applications and extensions. *Bull. Amer. Math. Soc.*, 63(2):199–331, 2026. doi:10.1090/bull/1880. ↑8, ↑9, ↑19

A Two laser methods

In this appendix, we discuss the intrinsic relation between the two kinds of laser methods: that of Strassen [Str87] and that of Coppersmith and Winograd [CW90]. We have long noticed that different communities in algebraic complexity theory tend to prefer different versions of the laser method, yet we have found no explicit discussion of this intrinsic connection in the literature, apart from a few brief remarks, e.g., [CFTZ22, Footnote 5]; we therefore provide a thorough explanation here.

We begin with the setting of our presentation. Consider a tensor T whose mode spaces are decomposed into subspaces $U = \bigoplus_{i\in I} U_i$, $V = \bigoplus_{j\in J} V_j$ and $W = \bigoplus_{k\in K} W_k$. Accordingly, we write T as a sum of its components

$$T = \sum_{i,j,k} T_{ijk}, \qquad T_{ijk} \in U_i \otimes V_j \otimes W_k,$$

and call the set $\Phi := \{(i,j,k) : T_{ijk} \neq 0\} \subseteq I \times J \times K$ the *blockwise support* of T.

Combinatorial framework. The goal of both laser methods is to find a combinatorial way to degenerate T into a direct sum of some of the components T_{ijk}. Being combinatorial, such a way depends only on the underlying support structure Φ.

The Coppersmith–Winograd laser method views Φ as a tripartite hypergraph, and seeks a large *induced matching*. That is, we pick subsets $I' \subseteq I$, $J' \subseteq J$ and $K' \subseteq K$ of the same size m, such that the restriction $\Phi \cap (I' \times J' \times K')$ is a perfect matching: it consists exactly of m triples $(i_t, j_t, k_t) \in \Phi$, $t = 1, \ldots, m$, where i_t, j_t, k_t run through I', J', K' respectively. In this case, zeroing out all subspaces not indexed by I', J', K' turns the tensor into a direct sum:

$$T \geqslant \bigoplus_{t=1}^{m} T_{i_t j_t k_t}.$$

Strassen's laser method instead seeks a *combinatorial degeneration*, which means finding *gradings* $\alpha\colon I \to \mathbb{Z}$, $\beta\colon J \to \mathbb{Z}$ and $\gamma\colon K \to \mathbb{Z}$ such that

- every $(i,j,k) \in \Phi$ satisfies $\alpha(i) + \beta(j) + \gamma(k) \geqslant 0$, and
- the triples $(i,j,k) \in \Phi$ attaining $\alpha(i) + \beta(j) + \gamma(k) = 0$ form a matching Φ_0.

In this case, we can scale each U_i by $\lambda^{\alpha(i)}$, and similarly each V_j by $\lambda^{\beta(j)}$ and each W_k by $\lambda^{\gamma(k)}$, where λ is a formal parameter. This gives

$$T \cong \sum_{i,j,k} \lambda^{\alpha(i)+\beta(j)+\gamma(k)} T_{ijk},$$

and letting $\lambda \to 0$ yields the degeneration

$$T \trianglerighteq \bigoplus_{(i,j,k)\in\Phi_0} T_{ijk}.$$

Tight sets. It turns out that both laser methods find the situation easy when the support Φ is *tight*, which means that one can find embeddings $I, J, K \subseteq \mathbb{Z}$ such that $i + j + k = 0$ for every

$(i,j,k) \in \Phi$. Moreover, both methods find that one should reduce the general case to arguably the "densest" tight case[4]: from now on, for a positive integer n, we let Φ_n denote the set

$$\Phi_n := \{(i,j,k) : i+j=k,\ 0 \leqslant i,j,k < n\}.$$

It is not hard to see that Φ_n is the support of the truncated polynomial multiplication tensor $\langle \mathbb{F}[t]/(t^n) \rangle$. Here Φ_n is tight, since negating the third coordinate gives the desired equation. In what follows, we work with the equation $i+j=k$ directly.

We now demonstrate how the two laser methods obtain large matchings from Φ_n.

Strassen's laser method. Strassen's laser method [Str87] obtains a matching through a combinatorial degeneration. To be more precise, Strassen's original argument only treated the case where Φ is the support of a matrix multiplication tensor, which arises as the outer structure of the Strassen tensor presented in Section 1.1. To our knowledge, the first presentations for general tight sets are due to the textbooks [BCS97, §15.8] and [Blä13, §9].

Since every triple of Φ_n already satisfies the linear equation $i+j=k$, a simple strategy to obtain a matching is to retain only those triples that satisfy a second linear equation. If this second hyperplane is not parallel to the first one, their intersection is a line, and with a suitable choice of the equation, the surviving triples form a matching. For simplicity, we take the line $i=j$.

This can be achieved by a monomial degeneration: we assign exponents to the coordinates in each mode so that the total exponent of each triple $(i,j,k) \in \Phi$ is exactly $(i-j)^2$. This is possible thanks to the factorization

$$\lambda^{(i-j)^2} = \underbrace{\lambda^{2i^2}}_{U_i} \cdot \underbrace{\lambda^{2j^2}}_{V_j} \cdot \underbrace{\lambda^{-k^2}}_{W_k}, \qquad k = i+j,$$

which follows from $(i-j)^2 = 2i^2 + 2j^2 - (i+j)^2$. In other words, $\alpha(i) = 2i^2$, $\beta(j) = 2j^2$ and $\gamma(k) = -k^2$ form a combinatorial degeneration of Φ: the total exponent $(i-j)^2$ is a square, so it is nonnegative and vanishes only when $i=j$, and the triples $(i,i,2i)$ clearly form a matching. For $\Phi = \Phi_n$, this tells us that

$$\Phi_n \text{ can be monomially degenerated to a matching of size } \lceil n/2 \rceil.$$

It is not hard to see that the above argument works for any second affine linear equation, by sieving with its square; this is also the key to the proof of Lemma 2.3.

Coppersmith–Winograd laser method. The Coppersmith–Winograd laser method [CW90] obtains an induced matching, by observing that the problem is almost already solved by a known combinatorial construction, namely *3-arithmetic-progression-free (3AP-free) sets*. We say that a subset $A \subseteq \mathbb{Z}$ is 3AP-free if every arithmetic progression inside it is trivial:

$$x,y,z \in A,\ x+z=2y \implies x=y=z.$$

One can clearly see that if Φ is the set of triples satisfying $i+k=2j$, then a 3AP-free set A can be directly applied to restrict all three coordinates, which yields an induced matching of size $|A|$. With a slight affine change of coordinates, the same works for any tight set; we show it

[4] Actual applications of the laser method need to deal with certain sparse tight sets, which require a careful *randomization* argument to reduce to the case we present here. Since this is technical and not related to our discussion, we omit the details.

for Φ_n. If A is a 3AP-free subset of $\{0, \dots, \lceil n/2 \rceil - 1\}$, then Φ_n has an induced matching of size $|A|$: we take

$$I' = J' = A, \qquad K' = 2 \cdot A := \{2a : a \in A\}.$$

Indeed, every surviving triple (i, j, k) gives a solution $a + a' = 2a''$ with $i = a$, $j = a'$ and $k = 2a''$ for some $a, a', a'' \in A$. Since A is 3AP-free, we must have $a = a' = a''$, so the surviving triples are exactly $(a, a, 2a)$ for $a \in A$, which form an induced matching.

It then remains to ask how large A can be, and this question turns out to be deep. First, it is impossible to retain a linear number of triples, as the combinatorial degeneration does: Roth [Rot53] first proved that $|A| = o(n)$ must be sublinear. After a long line of research, it was recently proved [KM23, BS23, Rag26] that we must have $|A| \leqslant n/2^{\Omega((\log n)^\beta)}$ for some $\beta > 0$. On the other hand, Behrend [Beh46] gave a construction almost matching this type of bound.

Theorem A.1 ([Beh46])**.** *There is a 3AP-free subset of $\{0, \dots, n-1\}$ of size $n/2^{O(\sqrt{\log n})}$.*

Combining Theorem A.1 with the above reduction, we obtain the following.

Corollary A.2. *The set Φ_n has an induced matching of size $n/2^{O(\sqrt{\log n})}$.*

A.1 From degenerations to zeroing outs

At first glance, there seems to be no relation between Strassen's combinatorial degeneration and Coppersmith and Winograd's use of 3AP-free sets. However, this is not the case.

In this subsection, we show a general way to turn a combinatorial degeneration into a zeroing out, at a cost that depends on the *order* of the degeneration. Readers familiar with tensor degenerations may recall the analogous argument that turns a border rank estimate into a rank estimate. We remark that this connection has already been noted in [CVZ23, Thm. 4.11] and [AV18b, Cor. 4.2 of the arXiv version].

We first rephrase the combinatorial degeneration so that its order can be defined. Instead of requiring the minimum of $\alpha(i) + \beta(j) + \gamma(k)$ to be zero, we add constants to α, β and γ, so that we may assume that they take nonnegative values. Then the minimum of $\alpha(i) + \beta(j) + \gamma(k)$ over Φ is some integer $h \geqslant 0$, which we call the *order* of the degeneration; the retained triples are exactly those with $\alpha(i) + \beta(j) + \gamma(k) = h$. In this case, we say that Φ combinatorially degenerates to the matching Φ_0 with order h.

Looking back at the argument for Strassen's laser method, we have the following.

Proposition A.3. *The set Φ_n combinatorially degenerates to a matching of size $\lceil n/2 \rceil$ with order $(n-1)^2$.*

Proof. We shift the gradings of Strassen's sieve to nonnegative values, by taking $\alpha(i) = 2i^2$, $\beta(j) = 2j^2$ and $\gamma(k) = (n-1)^2 - k^2$. Then the total exponent of each $(i, j, k) \in \Phi_n$ is $(i - j)^2 + (n-1)^2$, whose minimum $(n-1)^2$ is attained exactly by the $\lceil n/2 \rceil$ triples $(i, i, 2i)$. □

We now show how to turn a combinatorial degeneration directly into a zeroing out. The argument below is essentially the weight-layer argument of [AV18a, Lem. 6], which states an asymptotic version with an unspecified constant; here we make the dependence on h explicit.

Lemma A.4. *If Φ combinatorially degenerates to a matching Φ_0 of size D with order h, then Φ contains an induced matching of size at least $D/(h+1)^2$.*

Proof. The argument is simply an averaging principle. Each retained triple $(i, j, k) \in \Phi_0$ has a *degree triple* $(\alpha(i), \beta(j), \gamma(k))$, which consists of nonnegative integers summing to h. There are

at most $(h+1)^2$ such degree triples, so some degree triple (a,b,c) is shared by at least $D/(h+1)^2$ retained triples. We take

$$I' = \alpha^{-1}(a), \qquad J' = \beta^{-1}(b), \qquad K' = \gamma^{-1}(c),$$

that is, we zero out all coordinates in the three modes whose degrees are not a, b and c, respectively. Every triple $(i,j,k) \in \Phi \cap (I' \times J' \times K')$ satisfies $\alpha(i)+\beta(j)+\gamma(k) = a+b+c = h$, so it is a retained triple with degree triple (a,b,c). Conversely, every such retained triple survives. Since Φ_0 is a matching, these triples form an induced matching of the claimed size. □

From Strassen to Coppersmith–Winograd. For Strassen's laser method, this conversion is usually too expensive, since the order of the degeneration is large compared to D: for Φ_n, the order is $(n-1)^2$, while D is only $\lceil n/2 \rceil$. However, the high order becomes much less problematic once we take tensor powers. Indeed, the N-th power $\Phi^{\times N} \subseteq I^N \times J^N \times K^N$ is the blockwise support of $T^{\otimes N}$, and the gradings

$$\alpha^{(N)}(i_1,\dots,i_N) \coloneqq \alpha(i_1) + \cdots + \alpha(i_N),$$

and similarly $\beta^{(N)}$ and $\gamma^{(N)}$, show that $\Phi^{\times N}$ combinatorially degenerates to the matching $\Phi_0^{\times N}$ of size D^N with order Nh. By Lemma A.4, we obtain an induced matching of size at least

$$\frac{D^N}{(Nh+1)^2} = D^{N-o(N)}.$$

Thus, for sufficiently large N, we obtain a large induced matching by zeroing out coordinates, and asymptotically nothing is lost compared with the degeneration.

We now use this observation to show that the argument alone already recovers Corollary A.2, without using Theorem A.1.

Alternative proof of Corollary A.2. Applying Lemma A.4 to the N-th power of the degeneration in Proposition A.3, we see that $\Phi_n^{\times N}$ has an induced matching of size at least

$$\frac{\lceil n/2 \rceil^N}{(N(n-1)^2+1)^2} \geqslant \frac{(n/2)^N}{N^2 n^4}.$$

It remains to view $\Phi_n^{\times N}$ inside a single set $\Phi_{n'}$. We encode a vector $x \in \{0,\dots,n-1\}^N$ as the integer

$$\iota(x) \coloneqq \sum_{s=1}^{N} x_s (2n-1)^{s-1} < n' \coloneqq (2n-1)^N.$$

Since the sum of two digits is at most $2n-2$, no carries occur, so for $i,j,k \in \{0,\dots,n-1\}^N$ we have $\iota(i)+\iota(j) = \iota(k)$ if and only if $i+j=k$. In other words, after zeroing out all coordinates of $\Phi_{n'}$ outside the image of ι, we are left with exactly a copy of $\Phi_n^{\times N}$. Hence $\Phi_{n'}$ also has an induced matching of size at least

$$\frac{(n/2)^N}{N^2 n^4} \geqslant \frac{n'}{4^N N^2 n^4},$$

where we used $n' \leqslant (2n)^N$. Choosing $n \approx 2^{\sqrt{\log n'}}$ and $N \approx \sqrt{\log n'}$, the loss $4^N N^2 n^4$ is $2^{O(\sqrt{\log n'})}$, so $\Phi_{n'}$ has an induced matching of size $n'/2^{O(\sqrt{\log n'})}$. For a general n' that is not of the form $(2n-1)^N$, we restrict to the largest such value below it, which only loses a constant factor. □

We remark to unfamiliar readers that, if one tracks the above argument carefully, it in fact produces a 3AP-free subset of $\{0, \ldots, n-1\}$ of size $n/2^{O(\sqrt{\log n})}$, that is, it reproves Theorem A.1. Moreover, if one expands the argument, one finds that it is almost exactly Behrend's construction [Beh46]! That said, although the laser methods of Strassen and of Coppersmith and Winograd look very different, they rest on the same mechanism: Strassen's quadratic sieve, once expanded over tensor powers, is Behrend's construction in disguise.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il